HELSINGIN YLIOPISTO

Snowmobiles in the pants – Tekstiilialan tuotekuvausten neuroverkkokonekääntäminen

Saara Salminen
Syventävien opintojen tutkielma
Kääntämisen ja tulkkauksen maisteriohjelma
Humanistinen tiedekunta
Helsingin yliopisto
Toukokuu 2020

Tiedekunta – Fakultet – Faculty Humanistinen tiedekunta	Koulutusohjelma – Utbildningsprogram – Degree Programme Kääntämisen ja tulkkauksen maisteriohjelma	
Opintosuunta – Studieinriktning – Study Track Käännösteknologia		
Tekijä – Författare – Author Saara Salminen		
Työn nimi – Arbetets titel – Title Snowmobiles in the pants – Tekstiilialan tuotekuvausten neuroverkkokonekääntäminen		
Työn laji – Arbetets art – Level Syventävien opintojen tutkielma	Aika – Datum – Month and year Toukokuu 2020	Sivumäärä– Sidoantal – Number of pages 52

Tiivistelmä – Referat – Abstract

Tekstiili- ja muotiala elää monen muun alan tavoin murrosvaiheessa. Suomessa kehitetään entistä ympäristöystävällisempiä ja teknisempiä tekstiilimateriaaleja, joiden saaminen kansainvälisille markkinoille on monelle yritykselle elintärkeää.

Kansainvälistyminen tarkoittaa kuitenkin aina sitä, että tuotetietoa tarvitaan suomen lisäksi myös muilla kielillä. Kuluttaja haluaa lukea tuotetiedot omalla äidinkielellään, jotta hän ymmärtää tuotteen ominaisuudet ja pystyy vertailemaan tuotteita keskenään. Tuotekuvausten ja verkkosivujen kääntäminen useille kielille aiheuttaa verkkokauppiaille usein päänvaivaa, sillä monikieliselle tiedonhallinnalle ei välttämättä ole valmista prosessia. Monikielisen tiedonhallinnan prosessit vaihtelevat yrityksittäin, ja tuotetietojen kääntäminen on usein kallis ja monivaiheinen projekti, jonka lopputuloksena syntyvien käännösten laatu riippuu monesta tekijästä.

Tämän tutkielman tavoitteena on tutkia tekstiilialan tuotekuvauksia ja niiden sopivuutta neuroverkkokonekääntämiseen, sillä neuroverkkokonekääntämisestä voisi olla tulevaisuudessa hyötyä myös suomalaisten yritysten käännösprosesseissa. Tutkielmassa esikoulutettu geneerinen neuroverkkokonekäännin Marian uudelleenkoulutettiin, jotta saataisiin selville, parantaako uudelleenkoulutus konekäännösten laatua. Aineistona käytettiin pientä rinnakkaiskorpusta, joka koostuu tunnettujen ulkoiluvaateyritysten tuotekuvausteksteistä ja joka on koottu tätä tutkielmaa varten yritysten verkkosivuilta.

Tutkielman teoriaosuudessa esitellään tuotekuvauksia, niiden rakennetta sekä kielellisiä piirteitä ja tarkastellaan erilaisia tuotekuvauksia koskevia rajoituksia. Tuotekuvauksia tarkastellaan tutkielmassa erikoiskielenä ja vielä rajoitetummin minilektinä, jotta tarkastelu kattaisi useamman näkökulman. Lisäksi teoriaosuudessa esitellään neuroverkkokonekääntämistä ja sen tärkeimpiä teknologioita.

Tutkielman tuloksena voidaan todeta, että esikoulutetun geneerisen neuroverkkokääntimen voi adaptoida erikoisalalle myös hyvin pienellä määrällä koulutusmateriaalia niin, että BLEU-pisteet nousevat jo heti ensimmäisen koulutusajon jälkeen. Suurin parannus tapahtui erikoisalan termeissä, jotka neuroverkkokonekäännin oppi tehokkaasti pienestä koulutusaineistosta huolimatta. Koulutusaineisto aiheutti kuitenkin myös ongelmia: tuotekuvausten hyvin rajoittunut kieli ja lyhyet virkkeet heikensivät konekääntimen laatua ja aiheuttivat käännössegmenttien lyhentymistä ja poistoja. Ongelmaa ei saatu kokonaan korjattua, mutta laatua pystyttiin parantamaan toistamalla uudelleenkoulutus niin, että esikoulutusmateriaalissakin mukana olleita segmenttejä lisättiin uudelleenkoulutusaineistoon.

Tuotekuvaukset ja niiden neuroverkkokonekääntäminen tarjoavat runsaasti jatkotutkimusmahdollisuuksia, ja erikoisalakoulutetulle neuroverkkokonekääntimelle on varmasti kysyntää alati kasvavilla markkinoilla, kun käännöksiä toivotaan aikaisempaa nopeammilla aikatauluilla ja yhä edullisimmin kustannuksin.

Avainsanat – Nyckelord – Keywords
tuotekuvaus, neuroverkkokonekääntäminen, konekääntäminen, uudelleenkoulutus, käännösteknologia

Säilytyspaikka – Förvaringställe – Where deposited
Keskustakampuksen kirjasto

Muita tietoja – Övriga uppgifter – Additional information

Sisällysluettelo

Kuvaluettelo

Taulukkoluettelo

Kaavaluettelo

Lyhenneluettelo

FFN	Eteenpäin syöttävä neuroverkko (Feed-forward neural network)
LSTM	Pitkäkestoinen työmuisti (Long short-term memory)
RNN	Rekurrentti neuroverkko (Recurrent neural network)

1 Johdanto

Tekstiili- ja muotiala elää monen muun alan tavoin murrosvaiheessa. Suomessa kehitetään entistä ympäristöystävällisempiä ja teknisempiä tekstiilimateriaaleja, joiden saaminen kansainvälisille markkinoille on monelle yritykselle elintärkeää. Samalla käyttöön on otettu uusia kiertotalouden ja vastuullisuuden malleja.

Kansainvälistyminen tarkoittaa kuitenkin aina sitä, että tuotetietoa tarvitaan suomen lisäksi myös muilla kielillä. Kuluttaja haluaa lukea tuotetiedot omalla äidinkielellään, jotta hän ymmärtää tuotteen ominaisuudet ja pystyy vertailemaan tuotteita keskenään. Tuotekuvausten ja verkkosivujen kääntäminen useille kielille aiheuttaa verkkokauppiaille usein päänvaivaa, sillä monikieliselle tiedonhallinnalle ei välttämättä ole valmista prosessia. Monikielisen tiedonhallinnan prosessit vaihtelevat yrityksittäin, ja tuotetietojen kääntäminen on usein kallis ja monivaiheinen projekti, jonka lopputuloksena syntyvien käännösten laatu riippuu monesta tekijästä.

Monikielisen tiedonhallinnan lisäksi haasteita laadunhallintaan tuovat myös tuotekuvauksille asetetut vaatimukset ja odotukset. Tuotekuvauksissa yhdistyy eri käännöstrategioita vaativia elementtejä, sillä luovan otteen lisäksi kääntäjältä odotetaan tekstiilialan termistön tuntemusta. Virheelliset käännökset johtavat pahimmillaan kuluttajan harhaanjohtamiseen, kun esimerkiksi vettä hylkivää takkia markkinoidaan vedenpitävänä käännösvirheen vuoksi. Harhaanjohtavien tai totuudenvastaisten tietojen antaminen on kuluttajansuojalain mukaan kielletty markkinoinnissa, jos tiedot ovat omiaan johtamaan siihen, että kuluttaja tekee ostopäätöksen tai muun kulutushyödykkeeseen liittyvän päätöksen, jota hän ei ilman annettuja tietoja olisi tehnyt. Näin ollen virheet tuotekuvauksissa ja niiden käännöksissä saattavat johtaa reklamaatioihin ja lisäkuluihin, kun kuluttaja palauttaa vedenpitäväksi markkinoidun takin sen puutteellisten ominaisuuksien vuoksi. Omat haasteensa käännöksiin tuovat myös merkkirajoitukset ja monikieliseen hakukoneoptimointiin liittyvät säännöt ja ohjeet.

Tämän tutkielman tavoitteena on tutkia tekstiilialan tuotekuvauksia ja niiden sopivuutta neuroverkkokonekääntämiseen, sillä neuroverkkokonekääntämisestä voisi olla tulevaisuudessa hyötyä myös suomalaisten yritysten käännösprosesseissa. Tutkielmassa esikoulutettu geneerinen neuroverkkokonekäännin Marian uudelleenkoulutetaan erikoisalakohtaisella aineistolla, jotta

saataisiin selville, parantaako uudelleenkoulutus konekäännösten laatua. Aineistona käytetään pientä rinnakkaiskorpusta, joka koostuu tunnettujen ulkoiluvaateyritysten tuotekuvausteksteistä ja joka on koottu tätä tutkielmaa varten yritysten verkkosivuilta. Kriteerinä on se, että tekstien on löydyttävä yritysten verkkosivuilta sekä suomeksi että englanniksi. Tutkielmassa tullaan hyödyntämään materiaalia seuraavilta pohjoismaisilta brändeiltä: Luhta, Icepeak, Rukka, Torstai, Marimekko, Minna Parikka, Fjällraven, Haglöfs, Lovia, Didriksons, Joutsen ja Revolution Race.

Tutkielman teoriaosuudessa esitellään tuotekuvauksia, niiden rakennetta sekä kielellisiä piirteitä ja tarkastellaan erilaisia tuotekuvauksia koskevia rajoituksia. Tuotekuvauksia tarkastellaan tutkielmassa erikoiskielenä ja vielä rajoitetummin minilektinä, jotta tarkastelu kattaisi useamman näkökulman. Lisäksi teoriaosuudessa esitellään neuroverkkokonekääntämistä ja sen tärkeimpiä teknologioita.

Teoriaosuuden jälkeen Helsingin yliopiston kieliteknologian professorin Jörg Tiedemannin esikouluttama neuroverkkokonekäännin Marian koulutetaan yritysten verkkosivuilta kerättyjen kaksikielisten tuotekuvausten avulla. Osaa aineistosta käytetään koulutukseen ja osaa kääntämiseen, jotta koulutus- ja testiaineistona ei käytetä samaa tekstiä. Esikoulutetulla neuroverkkokonekääntimellä käännetään teksti kieliparissa suomi-englanti, jonka jälkeen käännösten laatua arvioidaan automaattisesti BLEU-pistein. Tämän jälkeen neuroverkkokäännin koulutetaan tuotekuvauksista koostuvalla aineistolla, ja käännösten laatua arvioidaan jälleen automaattisesti. Lopuksi käännösten tuloksia verrataan keskenään. Samalla todetaan, onko tutkielmassa käytettävä koulutusaineisto riittävän kattava parantaakseen esikoulutetun neuroverkkokonekääntimen laatua.

2 Teoreettinen viitekehys

Ennen konekääntimen koulutusta perehdytään tuotekuvausten sisältöön, rakenteeseen ja muihin kielellisiin piirteisiin, jotka erottavat tekstiilialan tuotekuvaukset muista tekstityypeistä. Tässä luvussa esitellään aineistoon kuuluvia tuotekuvauksia ja analysoidaan tuotekuvauksien rakennetta konekääntämisen näkökulmasta, jotta tutkielman tuloksia voitaisiin myöhemmin selittää teoriaosiossa esitetyillä tiedoilla. Tuotekuvauksia tarkastellaan tutkielmassa erikoiskielenä ja vielä rajoitetummin minilektinä, jotta tarkastelu kattaisi useamman näkökulman.

2.1 Erikoiskieli ja minilekti

Tuotekuvauksissa käytetyn kielen voidaan katsoa sisältävän useita kontrolloidun kielen piirteitä. Tuotekuvauksia ei kuitenkaan voida katsoa kontrolloiduksi kieleksi, sillä kontrolloitu kieli on keinotekoisesti tehtyä, vaikka se perustuukin luonnolliseen kieleen. Kontrolloitu kieli on luonnollisesta kielestä muokattu kielen variaatio, jota säätelevät erilaiset kieliopilliset (esim. lauseiden pituus, substantiivien määrä, passiivin käyttö) ja tyylilliset rajoitukset. Säännöillä pyritään vähentämään luonnollisen kielen monimerkityksisyyttä, joka vaikeuttaa konekääntämistä. (Kaji 1999: 37–38.)

Tuotekuvaukset voidaan näkökulmasta riippuen luokitella erikoiskieleksi tai minilektiksi, sillä ne sisältävät strukturoitua kieltä, jonka sanasto ja syntaksi ovat luonnollista kieltä rajoitetummat. Kontrolloidusta kielestä poiketen nämä rajoitukset ovat kuitenkin kehittyneet luonnollisesti tuotekuvauksiin kohdistuvien vaatimuksien takia. Tärkeimpiä rajoittavia tekijöitä tuotekuvauksissa ovat merkkimäärään ja termistöön sekä lauseiden yksinkertaisuuteen liittyvät rajoitukset. Näitä rajoituksia käsitellään tarkemmin luvussa 2.3.1.

2.1.1 Erikoiskieli

Suomen kieli – kuten harva muukaan kieli – on yhtenäinen kokonaisuus. Siitä voidaan erottaa pienempiä järjestelmiä tai osakieliä, jotka poikkeavat sanastoltaan ja virkerakenteeltaan yleiskielestä. Näitä yleensä tietyn tieteenalan tai ammatti- ja harrasteryhmien kielimuotoja kutsutaan erikoiskieliksi. (Haarala 1981: 9–10.) Suomalainen (2002) kirjoittaa erikoiskielten ja yleiskielen vuorovaikutussuhteesta seuraavasti:

> ” Erikoiskielet eivät ole yleiskielestä eriytyneitä kielimuotoja, vaan yleiskielen ja erikoiskielten välillä on vuorovaikutussuhde, joka elää koko ajan. Erikoiskielten termejä siirtyy yleiskieleen samalla kun yleiskielen sanoja otetaan erikoiskielten termeiksi. Erikoisalojen termeihin törmääkin yhä useammin myös arkielämässä esimerkiksi tiedotusvälineiden, erilaisten käyttöohjeiden tai vaikkapa televisiosarjojen kautta.”

Nykyään kuluttajat kohtaavat erikoiskieliä myös esimerkiksi verkkokaupassa asioidessaan, kun he lukevat erilaisten tuotteiden tuotekuvauksia tuotteiden ominaisuuksiin perehtyessään.

Erikoiskieliä voidaan näkökulmasta riippuen tarkastella teknolekteinä, kielen variaatioina, kielen variantteina tai jargoneina (Grygiel 2017: 3–4). Suppeimmillaan erikoiskieliä voidaan tarkastella vain termien käytön näkökulmasta, mutta laajemmasta näkökulmasta erikoiskieliin liittyy myös niissä käytetyn kielen rakenne tai tekstin yleinen muodostumistapa. Näin ollen tarkasteluun astuu myös erikoiskielten suhde kielijärjestelmään sekä teksti- ja tyylilajikysymykset. (Niemikorpi: 1996: 97). Tässä tutkielmassa tuotekuvauksia tarkastellaan ensin erikoiskielen näkökulmasta, jonka jälkeen esitellään erikoiskielen alakäsite minilekti.

Erikoiskieliä voidaan luonnehtia kielellä käsiteltävän tiedon avulla: yleiskieli sisältää yleistietoa ja erikoiskielet sisältävät erityistietoa. Erikoiskieltä sisältävät tekstit tunnistaa usein siitä, että niiden ymmärtäminen vaatii kyseessä olevan erikoisalan koulutusta. (Niemikorpi 1996: 100.) Erikoiskielet eroavat yleiskielestä niin sanastoltaan, morfosyntaksiltaan kuin rakenteeltaan, mutta tutkimuksissa on keskitytty etenkin erityiskielen sanastoon ja näin ollen erikoiskieliä on tarkasteltu pitkään lähinnä terminologian näkökulmasta (Grygiel 2017: 3–4).

Sanastoa ja erikoiskieltä kuvattaessa puhutaankin usein termeistä. Tässä tutkielmassa mukaillaan Laurénin & Nordmanin (1987: 81) määritelmää, jonka mukaan termi on tietyn erikoisalan käsitteen kielellinen ilmaus, jonka tunnistaminen vaatii erikoisalan tuntemusta. Tekstiilialan

termien tunnistamisessa on olennaista myös konteksti, jota käsitellään tarkemmin luvussa 2.3.4. Myös Haaralan (1981: 15-16) määritelmä vastaa tutkielmassa tarkasteltavan tekstiilialan erikoiskielen termejä. Haarala määrittelee termin seuraavien kolmen kriteerin avulla:

a) termi kuuluu nimenomaan erikoisalan kieleen
b) termin on oltava vakiintunut alan kielenkäytössä
c) termin täytyy olla tietyn tarkasti määritellyn käsitteen nimitys.

Erikoiskielen termit saattavat syntyä yleiskielen sanan pohjalta, mutta saman teknolektin termiä ei mielellään käytetä useissa merkityksissä eikä yhteen tiettyyn merkitykseen tule viitata useilla eri termeillä (Laurén & Nordman 1987: 79–80). Termit eroavat yleiskielen sanoista siis juuri sopimuksenvaraisten merkitystensä takia.

Tekstiilialan tuotekuvauksissa esiintyy yleiskielelle tyypillisiä sanoja, jotka ovat tuttuja jokaiselle maallikolle. Niiden voidaan kuitenkin katsoa kuuluvan erikoiskieliseen sanastoon, sillä niiden merkitys on tuotekuvauksissa eri kuin yleiskielessä. Tällaisia sanoja esiintyy tuotekuvauksissa runsaasti, mutta esimerkiksi sanat *vuori*, *tuuletusaukko* ja *sauma* (esimerkki 1 a ja b) esiintyvät usein sekä yleiskielessä että tekstiilialan erikoiskielessä. Erikoiskielessä niiden merkitys on kuitenkin erilainen kuin yleiskielessä.

(1) a) Siinä on lämmin teddyvuori, vedenpitävä pintamateriaali ja teipatut saumat.
b) Takissa on normaali istuvuus ja tuuletusaukko hartian takaosassa.

Erikoiskielisiä ilmauksia siirtyy yleiskieleen runsaasti, mutta samalla niiden erikoiskielinen merkitys usein hämärtyy, sillä maallikko ymmärtää käsitteen taustalla olevan ilmiön huomattavasti rajallisemmin kuin alan asiantuntija. Tämä johtaa siihen, että erikoiskielisen käsitteen sisältö yksinkertaistuu ja saa yleiskielisen merkityksen. (Suomalainen 2002.) Tämä yleiskielinen merkitys ei kuitenkaan aina vastaa alkuperäistä erikoiskielistä merkitystä. Tekstiilialalla hyviä esimerkkejä tällaisesta ilmausten siirtymisestä ovat **vedenpitävä** ja **vettähylkivä**. Maallikko ei välttämättä tunne ilmaisuja ja käyttää niitä esimerkiksi synonyymeina toisilleen, vaikka tekstiilialalla kyseessä on kaksi täysin erilaista valmistustekniikkaa. Vedenpitävyydessä tulee ottaa huomioon, että vaikka tuotteessa käytetty kangas olisi vedenpitävä, tulee lisäksi sen kaikkien saumojen olla teipattu, jotta tuote itsessään olisi vedenpitävä. Vedenpitävyys perustuu

pintamateriaalin alle laminoituun vedenpitävään kalvoon, kun taas vedenhylkivyys perustuu yleensä erityiseen pintakäsittelyyn. Tuotteen vedenpitävyyttä kuvataan vesipilariarvolla: 10 000 mm on erinomainen ja 5 000 mm hyvä vedenpitävyys. (Icepeak 2019.) Tätä harva maallikko ottaa huomioon vedenpitävästä takista puhuessaan, ja tekstiilialan erikoiskielen käsite yksinkertaistuu.

Kun vastaanottajana on ammattilaisen sijaan maallikko, vaikuttaa se voimakkaasti käytettyyn kieleen (Nordman 1989: 72). Erikoisaloilla, joiden termistöä käytetään laajasti myös erikoisalan ulkopuolella, on erityisen tärkeää, että alan termistö on ymmärrettävää myös muille kuin alan asiantuntijoille. Tästä syystä erikoiskielen termien tulee olla tarkkaan harkittuja ja omakielisiä. (Suomalainen 2002.) Lisäksi tietyn erikoisalan asiantuntijoiden tulisi käyttää samoja nimityksiä ja termejä, jotta yhtenäisyys säilyisi. Yhtenäisyyden säilyminen vaatii kuitenkin sen, että yhtenäisyys on ensin saavutettu. Tekstiilialalla yhtenäisyyteen on vielä matkaa: vaihtelua esiintyy etenkin uusissa teknologioissa, sillä yritykset pyrkivät erilaistamaan tuotteensa muista vastaavista kehittämällä jatkuvasti uusia nimityksiä eikä alalla ole selkeitä yhtenäistämiskäytäntöjä käytetylle sanastolle. Yhtenäisyyden saavutus ja säilyttäminen vaatii tietoista ja jatkuvaa sanastotyötä (Haarala 1981: 12), mutta yrityksissä pääpaino on myyntiä edistävässä sisällöntuotannossa, eikä yhtenäisyyteen panosteta. Kielen hallinnointi on erityisen hankalaa tekstiilialalla, jossa uusia sanoja ilmestyy jatkuvasti ja kielenkäytön normit luodaan lähes aina käytännön kautta.

2.1.2 Minilekti

Tässä luvussa tarkastellaan Nordmanin (1994) esittelemää minilektin käsitettä ja perehdytään minilektien määritelmään, syntyyn ja kriteereihin. Tuotekuvausten tarkastelu minilektinä on tutkielman aihepiirin kannalta tärkeää, sillä minilektien rajallinen sanasto ja kielioppi ovat herättäneet kiinnostusta myös konekääntämisen alalla (Laurén & Nordman 1987: 59, ks. myös Lehrberger 1986: 81–106; Seljan 2000). Neuroverkkoihin liittyen vastaavaa tutkimusta ei ole kuitenkaan vielä juurikaan tehty.

Suurin osa erikoiskielistä muodostuu teknolekteistä. Hyvänä esimerkkinä teknolektistä toimii lakitieteen teknolekti. (Laurén & Nordman 1987: 33.) Kun teknolektin kieltä käyttää vain hyvin rajoitettu määrä ihmisiä – tai kuten tuotekuvauksissa – kun teknolekti liittyy erittäin rajoitettuun erikoisalueeseen kuten tekstiilialaan, on kyseessä minilekti (Nordman 1994: 23).

Minilektit esiintyvät siis teknolektin sisällä ja ovat näin teknolektin alakäsite, samalla tavalla kuin teknolektit ovat erikoiskielen alakäsite. Usein esimerkiksi harrastuksiin liittyvillä minilekteillä, kuten neulomisohjeilla tai ravivihjeillä, ei kuitenkaan katsota olevan erillistä teknolektiä, koska teknolektillä viitataan usein juuri tieteelliseen kieleen (Nordman 1994: 10-11). Esimerkiksi ruokaohjeita ja sääennusteita voidaan pitää minilekteinä. (Nordman 1994: 30).

Kuva 1. Erikoiskielen, teknolektin ja minilektin käsitejärjestelmä (Nordman 1994: 11).

Minilektit ovat erillisiä kielellisiä alajärjestelmiä, joita säätelevät erilaiset kriteerit ja rajoitukset. Nämä kriteerit ja rajoitukset liittyvät tekstien semanttisiin, leksikaalisiin ja syntaktisiin piirteisiin, joita esitellään tuotekuvausten osalta tarkemmin luvun 2.3 alaluvuissa. Minilektien on oltava semanttisesti rajoittuneita ja niiden termien on oltava vakiintuneita. Samalla minilektiin kuuluvien tekstien tulee olla semanttisesti, syntaktisesti ja rakenteellisesti yhtenäisiä. Rajoitukset ja yhtäläisyydet johtavat siihen, että minilektien tekstit ovat usein lyhyempiä, yksinkertaisempia ja kaavamaisia kuin teknolektien tekstit ja näin ollen ne ovat myös vähemmän monitulkintaisia. (Laurén & Nordman 1987: 50–53.) Lisäksi uusien tekstien laadinnassa käytetään vanhoja malleja (Nordman 1994: 38–39), eli minilektiin kuuluvat tekstit säilyttävät yksinkertaisuutensa ja kaavamaisuutensa myös eri kirjoittajien ja esimerkiksi uusien tuotteiden kohdalla.

Minilektien teksteissä viestin on oltava tiivis ja välityttävä nopeasti (Laurén & Nordman 1987: 53). Tästä syystä minilekteissä esiintyvät lausetyypit ovat hyvin rajallisia ja tiiviin esittämismuodon mahdollistavia lausetyyppejä suositaan. Eri minilekteissä painotetaan erilaisia keinoja: imperatiivin käyttö on suosittua erilaisissa ohjeissa, joissa pyritään minilekteille tyypilliseen suoraan ja tiedottavaan ilmaisuun. Keskusteleva ote ja tunteisiin vetoaminen ovat harvinaisempia minilekteissä. (Nordman 1994: 41.) Poikkeuksena voidaan kuitenkin pitää tuotekuvauksia, joissa suoran ja tiedottavan ilmaisun lisäksi pyritään vetoamaan lukijan tunteisiin

ja luomaan vaikutelmaa luonnollisesta keskustelusta. Tuotekuvausten tieto-/suostuttelusuhteen määrää esitellään tarkemmin luvussa 2.2.2.

Rajallisten lausetyyppien lisäksi minilekteissä keskitytään usein hyvin rajattuun määrään tietyn sanaluokan sanoja. Näin ollen monen minilektin sanasto on huomattavasti yleiskielen sanastoa suppeampaa. Minilekteissä esiintyy usein tiettyjä stereotyyppisiä substantiiveja ja verbejä, jotka liittyvät minilektin rajoitettuun sisältöön (Nordman 1994: 43). Tekstiilialan tuotekuvauksissa tällaisia substantiiveja ovat esimerkiksi erilaiset materiaalit, kuten *neulospinta, untuva, kumipohja* tai *luonnonkuitu* ja tuotteiden osat, kuten *hihansuu, helma, vuori* tai *kaulus*. Verbeissä erityisen tyypillisiä ovat positiivisen konnotaation omaavat indikatiivimuodot, kuten *parantaa, hengittää, lämmittää, suojaa* ja *korostaa*. Lisäksi tuotekuvauksissa käytetään runsaasti tuotteen materiaalin ominaisuuksiin viittaavia adjektiiveja, kuten *lämmin, kestävä, kevyt, hengittävä, tyylikäs, joustava* ja *pehmeä*. Joissakin minilekteissä adjektiiveja ei kuitenkaan käytetä olleenkaan (Nordman 1994: 43).

2.2 Tekstiilialan tuotekuvaukset verkkokaupoissa

Internet on tuonut mukanaan täysin uuden tavan tehdä kauppaa. Näin myös perinteisen ostoprosessin vaiheet ovat murroksessa. Kun kaikki tiedon etsinnästä aina tuotteen tilaamiseen voidaan tehdä verkossa, on myös ostoprosessi nopeutunut. Perinteisessä ostoprosessissa voidaan katsoa olevan viisi vaihetta (Kotler & Keller 2006: 191), jotka Isohookana (2007: 80) on suomentanut seuraavasti:

1. ongelman tai tarpeen määrittäminen
2. tiedon hankinta
3. vaihtoehtojen arviointi
4. ostopäätös
5. ostopäätöksen jälkeinen käyttäytyminen

Internetin myötä ostoprosessin ei kuitenkaan tarvitse enää olla aikaa vievä toiminto, joka etenee vaihe vaiheelta, sillä kaupankäynti verkossa mahdollistaa ostoksenteon eri vaiheiden yhdistelyn ja uudelleenjärjestelyn. Näin ollen myös kuluttajien ostokäyttäytyminen on muuttunut ja he omaksuvat erilaisia tapoja hyödyntää internetiä ostoksia tehdessään. Tuotteista etsitään tietoja ja arvioita ja tuotteiden ominaisuuksia, laatua ja hintaa vertaillaan ennen ostopäätöksen tekemistä. Markkinoijan onkin tunnettava kuluttajan tiedontarve ostoprosessin eri vaiheissa mahdollisimman hyvin, jotta tiedetään, mitä ja miten tuotteista viestitään (Isohookana 2007: 80).

Ostoprosessin murroksen myötä myös tuotekuvausten merkitys on kasvanut. Kuluttajatuotteisiin liittyvä tieto on yhä monitahoisempaa, ja asiakkaan tarve tuotetiedolle ostohetkellä on kasvanut voimakkaasti (Janzen & Maass 2008). Tuotekuvausten lisäksi tuotteeseen liittyvät palvelu- ja tukitoiminnot sekä erilaiset arvostelut ja tuotteen linkitys yhteensopiviin tai vaihtoehtoisiin tuotteisiin ovat ostopäätöksen kannalta merkittävää tietoa kuluttajalle.

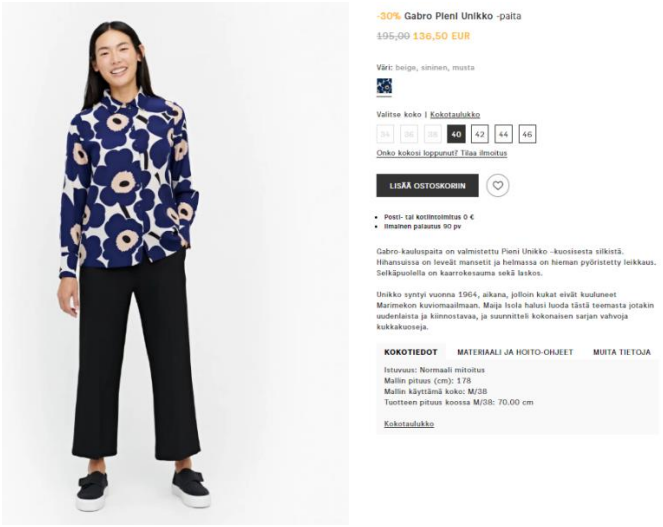
Verkkokaupassa tuotteen koskettaminen, sovittaminen tai muu tarkastelu ennen ostopäätöksen tekemistä on harvoin mahdollista eikä ostotilanteessa ole läsnä kivijalkaliikkeelle tyypillistä myyjää, joka voisi antaa lisätietoja tuotteen materiaaleista tai ominaisuuksista. Kuluttajan ainoa tiedonlähde on tällöin verkosta löytyvä tuotekuvaus ja muu tuotteeseen liittyvä tieto sekä mahdolliset muiden kuluttajien jättämät arvostelut. Näin ollen tuotekuvauksilla on hyvin suuri vaikutus kuluttajien ostokäyttäytymiseen.

2.2.1. Tuotekuvauksen määritelmä

Vuokon (2003: 17) mukaan markkinointiviestintään kuuluvat kaikki ne viestinnän elementit, joiden tarkoituksena on luoda yrityksen ja sen eri sidosryhmien välillä vuorovaikutusta, joka johtaa lopulta markkinoinnin tuloksellisuuteen. Markkinointiviestinnällä pyritään siis vaikuttamaan joko suoraan tai välillisesti tuotteiden tai palveluiden myyntiin. Kun tämä viestintä tapahtuu digitaalisten kanavien, kuten internetin, sähköpostin tai matkapuhelinten kautta, puhutaan digitaalisesta markkinointiviestinnästä (Merisavo 2008: 19–20).

Tuotekuvaukset ovat osa digitaalista markkinointiviestintää. Tuotekuvauksella tarkoitetaan tässä tutkielmassa verkkokaupassa olevaa kirjoitettua kuvausta tuotteesta, sen tyypistä, ominaisuuksista ja sen soveltuvuudesta erilaisiin käyttötarkoituksiin.

Kuvaus löytyy verkkokaupasta vaatteen tuotesivulta. Tuotesivulta löytyy valokuvan lisäksi lyhyt kuvaus tuotteesta, listaus tuotteen ominaisuuksista ja käytetyistä materiaaleista ja materiaaliteknologioista sekä tieto koko- ja värivaihtoehdoista ilmaistuna joko tuotteen nimen yhteydessä tai kuvauksen lopussa verkkokaupasta riippuen. Tuotesivulta löytyy myös vaatteen hoito-ohje. Tuotesivulla voidaan esitellä myös valittuun tuotteeseen sopivia tai saman tuoteperheen vaatteita, sillä tuoteperheen korostus helpottaa uuden tuotteen lanseerausta (Isohookana 2007: 50).



Kuva 2. Esimerkki tuotesivusta.

2.2.2 Tuotekuvausten sisältö

Markkinointiviestinnän avulla kuluttajille luodaan odotuksia ja lupauksia yrityksen tuotteista ja palveluista (Isohookana 2007: 17). Tuotekuva on kuluttajan mielikuva konkreettisesta tuotteesta, johon tuotekuvaus pyrkii vaikuttamaan. Tuotekuvaukset ovat nimensä mukaisesti kuvauksia tuotteista: niiden avulla annetaan tietoa tuotteesta ja sen toiminnasta. Tuotteiden toimintaa ja ominaisuuksia käsitellään tuotekuvauksissa usein hyvin perusteellisesti, jotta varmistetaan, että kuluttaja saa mahdollisimman tarkkaa ja yksityiskohtaista tietoa siitä, mihin tarkoitukseen tuotetta käytetään ja mihin tuotteen toiminta perustuu.

Tuotteen viestintä ja erilaistaminen kilpailijoista lähtee tuotteen ja sen ominaisuuksien määrittelystä (Isohookana 2007: 50). Tuotekehityksen ja markkinoinnin yhteistyö on välttämätöntä, jotta esimerkiksi vaatesuunnittelijan tekemät ratkaisut materiaalien ja teknologioiden suhteen saadaan tuotua esiin myös tuotekuvauksissa. Tuotekehityksen tuntemus tekniikasta yhdistettynä markkinoinnin ja tuotekuvauksia kirjoittavan sisällöntuottajan tuntemukseen markkinoista, kuluttajista ja käytettävistä termeistä takaavat yksityiskohtaiset tuotekuvaukset, jolloin myös erilaistaminen on helpompaa.

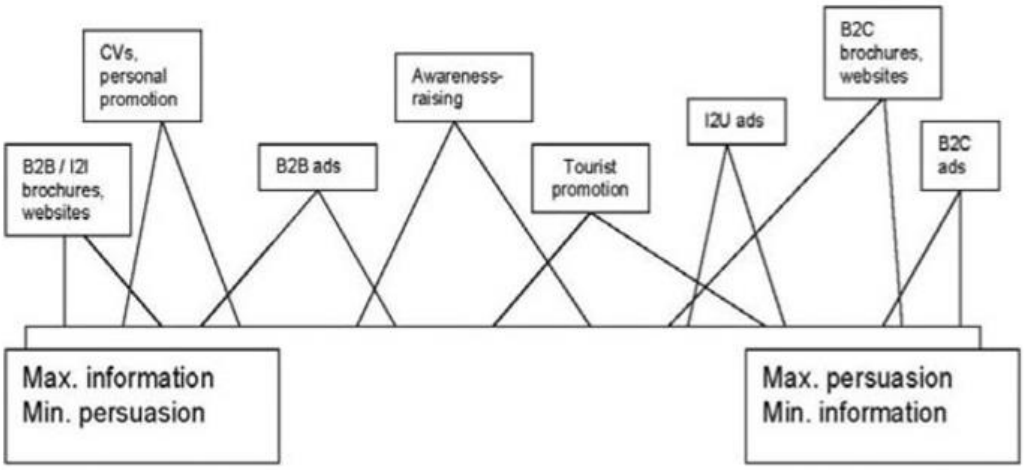
Tekstien voidaan ajatella koostuvan funktionaalisista jaksoista (Kuikka 2007: 43). Näin ollen myös tuotekuvaus rakentuu osista, joilla on jokaisella oma funktionsa kuluttajan aktivoinnissa ja tuotteen myymisessä. Aineistossa esiintyvät tuotekuvaukset koostuvat seuraavista funktionaalisista jaksoista, jotka Kuikka (2007: 57) havaitsi lehtimainosta analysoidessaan:

1) huomion herättäminen
2) käyttötarkoitukseen johdattelu
3) tuoteryhmän määrittely
4) tuotteen määrittely
5) kohderyhmän määrittely
6) eksplisiittinen pääväite eli myyntiväittämä
7) lisäväitteet
8) perustelu.

Jaksot eivät kuitenkaan ole lineaarisia, sillä jaksoa täydentäviä elementtejä (verbaaleja ja visuaalisia) on useissa eri mainoksen osissa: jaksot täydentävät toisiaan ja kokonaisuus muodostuu

eri elementtien yhteisvaikutuksesta. (Kuikka 2007: 53.) Tuotekuvauksissa eri elementtien yhteisvaikutus korostuu, sillä tekstillä pyritään täydentämään kuvan välittämää informaatiota ja tilan puutteen vuoksi kaikki kuvasta havaittava jätetään yleensä mainitsematta. Lisäksi tuotekuvauksille asetetut rajoitukset pakottavat yhdistelemään funktionaalisia jaksoja jopa yhden virkkeen sisällä.

Funktionaalisten jaksojen lisäksi tuotekuvauksia voidaan tarkastella myös niiden sisältämien piirteiden avulla. Tärkeiden ominaisuuksien listaamisen ohella lukijan kiinnostusta pyritään herättämään erilaisin kielellisin keinoin. Torresi (2014) on tutkinut mainosten ja muiden funktioltaan myyvien tekstien kääntämistä ja todennut, että kulutushyödykkeiden markkinoinnissa (engl. *business-to-consumer marketing, b-to-c, B2C*) tekstit sisältävät aina kahdenlaisia piirteitä: teknisiä ankkureita (engl. *technical anchors*) ja myynninedistämiseen ja vakuutteluun liittyviä elementtejä (engl. *boost elements*). Kyse on tieto-/suostuttelusuhteesta (engl. *information-to-persuasion ratio*), jossa eri genret sijoittuvat eri asteikolle niiden sisältämien elementtien perusteella. Teknisimmätkin tekstit voivat sisältää myynninedistämiseen ja suostutteluun liittyviä elementtejä, samalla tavalla kuin hyvin luovat ja myyvät tekstit saattavat sisältää teknisiä faktoja. (Torresi 2014: 27.)



Kuva 3. Erilaisten mainostekstilajien sijoittuminen tieto-/suostuttelusuhteen asteikolle. (Torresi 2014: 28).

Tuotekuvaukset sijoittuvat Torresin asteikolla kulutushyödykkeiden markkinoinnin osioon, sillä osio sisältää erilaiset esitteet ja verkkosivut. Näin ollen tuotekuvaukset sijoittuvat asteikolla niihin tekstilajeihin, jotka perustuvat enemmän myynninedistämiseen ja suostutteluun liittyviin elementteihin kuin teknisiin tietoihin. Kuvassa 3 näkyvä suurehko jakauma kulutushyödykkeiden markkinoinnin tiedon ja vakuuttamisen suhteessa on tekstiilialan tuotekuvauksien näkökulmasta realistinen, sillä tuotekuvauksissa pääpaino voi vaihdella teknisen tiedon ja vakuuttelevan tekstin välillä hyvinkin paljon riippuen siitä, onko kyseessä erilaisin ominaisuuksin varusteltu ulkoilutakki vai yksinkertainen polyesteripaita. Myös brändillä on suuri rooli tiedon ja vakuuttamisen suhteessa, sillä sisällöntuotantoa ohjaa brändin halu näyttäytyä tietynlaisena: koko perheen ulkoilutuotteita kuvataan tekstissä eri tavalla kuin pehmeää joogamallistoa tai kovaan yksilösuorittamiseen tarkoitettua tuoteperhettä.

Lisäarvon tuominen ja erilaistaminen ovat hyvä esimerkki tieto- ja faktapainotteisista tuotekuvauksista, jotka sisältävät useita teknisiä ankkureita. Tuotekuvauksissa pyritään tuomaan lisäarvoa tuotteelle eli listaamaan sellaisia tuotteen ominaisuuksia, jotka eivät näy tuotekuvassa tai ilmene muuten ominaisuusluettelossa. Tällaisia ominaisuuksia ovat etenkin tuotteessa käytetyt erikoismateriaalit ja teknologiat (esim. Dri-FIT, GORE-TEX). Lisäarvoa pyritään tuomaan myös antamalla vinkkejä yhteensopivista asusteista tai esittämällä tuote osana kokonaisuutta.

Erilaistamisessa esille tuodaan puolestaan ne tekijät, jotka erottavat tuotteen muista vastaavista tuotteista. Nykyään tällaisia argumentteja ovat yleensä etenkin eettisesti tuotetut materiaalit ja ympäristöystävälliset valmistustekniikat, kuten FairTrade-puuvilla, kierrätetty polyesteri tai mulesing-vapaa merinovilla. Ympäristöystävällisyyttä ilmaistaan usein erilaisin nonverbaalisin keinoin kuten kuvituksella (symbolit ja logot) tuotekuvan tai -nimen yhteydessä, mutta tuotekuvauksessa eettisyyttä päästään esittelemään tarkemmin:

(2) a) North Outdoorin käyttämä merinovilla on eettisesti tuotettua mulesing-vapaata villaa.

b) Kierrätetystä puuvillasta valmistetut tuotteet tulevat Pure Wastelta. Pure Waste on suomalainen tekstiiliteollisuuden yritys, joka tuottaa ekologisesti kestäviä lankoja, kankaita sekä vaatteita käyttäen vain materiaaleja, jotka muuten menisivät tuhottaviksi.

Vähemmän tietoa sisältävissä tuotekuvauksissa pääpaino on tiedon sijaan kuluttajaan vaikuttamisessa. Vaikutuskeinoja ja strategioita on monia, mutta Torresi (2014: 121–128) puhuu mainostekstilajien yhteydessä etenkin luovasta kielestä ja tunteita herättävästä kielestä (eng.

creative language ja *emotional language*). Luova kieli sisältää metaforia, sanaleikkejä, uudissanoja, allitteraatioita ja onomatopoeesia. Myös toisto ja intertekstuaalisuus ovat tärkeä osa luovaa kieltä. (Torresi 2014: 121–124.) Tunteita herättävä kieli (esimerkki c) sisältää paljon joko selkeästi positiivisia tai negatiivisia konnotaatioita omaavia termejä sekä sanoja, joissa on selkeä tunteellinen lataus (esim. *dream*, *fantastic*, *magic*). Tuotteen sijaan vaikuttamisen pääpaino voi olla myös kuluttajassa, jolloin puhuttelussa käytetään tyypillisesti sinuttelua ja monikon ensimmäistä persoonaa, kuten esimerkissä d. Puhuttelun lisäksi englannissa possessiivipronominia (*mine*, *yours*) ja suomessa persoonapronominien genetiivejä (minun, sinun) tai vaihtoehtoisesti substantiiviin liitettyä possessiivisuffiksia käytetään luonnollisen keskustelun vaikutelman luomiseksi. (Torresi 2014: 128.)

> c) Korkokenkä, jonka korko on klassinen punainen tikkari, syntyi herkullisesta ajatuksesta ja halusta nauttia jotain makeaa – niin usein kuin haluaa.
> d) On sunny early autumn days, all you need is a lightweight top and shorts for running.

Tiedon ja suostuttelun suhde on kiinnostava tutkimuskohde: voittaako tunne informaation vai haluaako kuluttaja tehdä ostopäätöksensä mieluummin pelkkiin faktoihin perustuen? Informatiiviset keinot ovat tehokkaita silloin, kun ne kiinnostavat kuluttajaa (Mustonen 2001: 45). Etenkin tekstiilialalla monet tuotteet ovat laadultaan hyvin samanlaisia, jolloin informatiivinen mainonta toimii yhä huonommin kuluttajiin. Tällöin positiivisiin tunteisiin vaikuttavat keinot suosittelevat paremmin kuin informaatioon perustuvat keinot. Mainonnan – ja samalla myös tuotekuvausten – sisältämälle informaatiolle on yhä vähemmän kysyntää, sillä kuluttajat voivat hakea tietoa helposti ja nopeasti muistakin lähteistä. Näin ollen kuluttajan huomion herättäminen on tarjottavaa informaatiota tärkeämpää (Malmelin 2003: 48). Kuvassa 3 näkyvä suurehko jakauma kulutushyödykkeiden markkinoinnin tieto-/suostuttelusuhteessa johtuneekin nimenomaan jatkuvasta arvioinnista sen suhteen, kuinka paljon tietoa ja tunnetta tarvitaan, jotta kuluttaja huomioi tuotekuvauksen ja vakuuttuu sen sisällöstä.

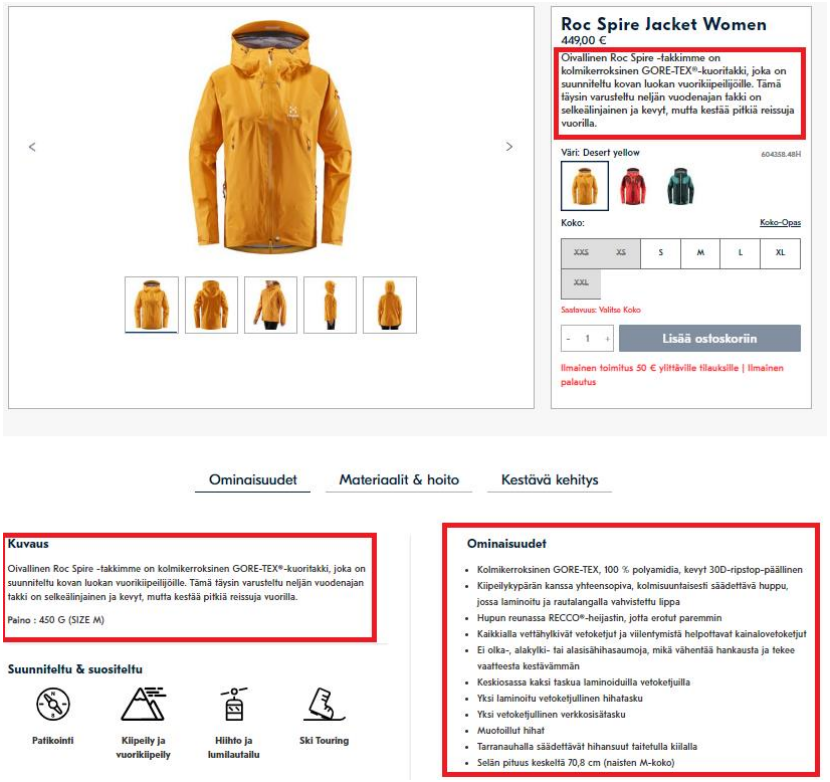
2.2.3 Tuotekuvausten rakenne ja kielelliset piirteet

Yhdessä tekstissä voi esiintyä kahden tai useamman tekstityypin piirteitä (Reiss & Vermeer 1986, 116–117.) Tuotekuvaukset yhdistelevät piirteitä niin informatiivisesta, ekspressiivisestä kuin operatiivisestakin tekstityypistä, sillä ne pyrkivät samanaikaisesti tarjoamaan tietoa, herättämään

positiivisia mielikuvia ja assosiaatioita sekä johdattelemaan kuluttajaa tekemään ostopäätöksen. Tuotekuvauksissa annettu informaatio on kuitenkin aina valikoitua, sillä tuotekuvausten ensisijainen funktio on mainostaa tuotetta.

Tuotekuvauksissa käytetyn kielen voidaan katsoa sisältävän useita kontrolloidun kielen piirteitä. Tuotekuvauksia ei kuitenkaan voida katsoa kontrolloiduksi kieleksi, sillä kontrolloitu kieli on keinotekoisesti tehtyä, vaikka se perustuukin luonnolliseen kieleen. Tuotekuvaukset voidaan luokitella erikoiskieleksi, sillä erityisen termistönsä lisäksi ne sisältävät tiettyjä tyylillisiä ja syntaktisia piirteitä, joita yleiskieli ei sisällä. Toisaalta tuotekuvauksien on oltava mahdollisimman selkeitä ja ymmärrettäviä, jotta niiden mainostamisfunktio säilyy. Kun erikoisalaa selitetään maallikolle, on asioita kuitenkin pakko yksinkertaistaa – samalla käytetty kieli lähestyy yleiskieltä (Laurén & Nordman 1987). Näin ollen tuotekuvausten määrittely ei ole yhtä suoraviivaista kuin monen muun tekstityypin määrittely.

Tuotekuvaukset on yleensä jaettu kahteen osaan, joista toinen koostuu juoksevasta tekstistä ja toinen ominaisuusluettelosta tai listasta (ks. kuva 4, osiot merkitty punaisella). Tekstiosio koostuu tuotteen nimestä ja virkefragmenttimuotoisesta tuotekuvauksesta. Ominaisuusluettelo sisältää usein samaa tietoa kuin tuotekuvaus, mutta materiaalien ja teknologioiden lisäksi ominaisuusluetteloon sisältyy usein tietoa, jota kuvauksessa ei mainita. Ominaisuusluettelossa mainitaan usein esimerkiksi tuotenumero ja muut värivaihtoehdot. (Jaaranen 2000: 14–15.)



Kuva 4. Esimerkki tuotekuvauksesta ja ominaisuusluettelosta.

2.3 Tuotekuvausten kääntäminen

Kansainvälistyminen tarkoittaa aina sitä, että tuotetietoa tarvitaan suomen lisäksi myös muilla kielillä. Kuluttaja haluaa lukea tuotetiedot omalla äidinkielellään, jotta hän ymmärtää tuotteen ominaisuudet ja pystyy vertailemaan tuotteita keskenään (Laenen & Moes 2019: 46). Tuotekuvausten ja verkkosivujen kääntäminen useille kielille aiheuttaa verkkokauppiaille usein päänvaivaa, sillä monikieliselle tiedonhallinnalle ei välttämättä ole valmista prosessia. Monikielisen tiedonhallinnan prosessit vaihtelevat yrityksittäin, ja tuotetietojen kääntäminen on usein kallis ja monivaiheinen projekti, jonka lopputuloksena syntyvien käännösten laatu riippuu monesta tekijästä. Tässä luvussa tarkastellaan lyhyesti tuotekuvausten kääntämiseen liittyviä haasteita, jonka jälkeen perustellaan, miksi juuri tuotekuvaukset sopisivat hyvin konekäännettäviksi.

Edellisessä luvussa esitelty tiedon ja vaikuttamisen suhde on avainasemassa tuotekuvausten kääntämisessä. Kaikista informatiivisimmatkin tekstit voivat sisältää suostuttelun keinoja ja päinvastoin: suostutteluun perustuvat tekstit voivat sisältää faktoja ja esimerkiksi numeerista dataa. Käännösstrategian valinta ei siis ole yksiselitteistä, sillä vaihtelu tiedon ja vaikuttamisen välillä voi tapahtua tekstin tai kappaleen lisäksi jopa lausetasolla. (Torresi 2014: 27.) Kun kyseessä on kuluttajaan vetoaminen, kuten luovan ja tunteita herättävän kielen käyttö, tulee käännöksen herättää kuluttajassa samanlaisia tunteita kuin lähtötekstin. Tarkkuutta tärkeämpää on siis käännöksen toimivuus. Teknisten ankkureiden kohdalla tilanne on päinvastainen, sillä käännösten on oltava mahdollisimman tarkkoja. (Torresi 2014: 69.) Tuotekuvauksissa tämä korostuu, sillä virheet teknisten ankkureiden kääntämisessä saattavat johtaa kuluttajan harhaanjohtamiseen, samalla kun virheet suostutteluun liittyvien elementtien kääntämisessä johtavat siihen, että tuotekuvaus menettää suostuttelevan funktionsa. Monet tuotekuvaukset sisältävät kulttuurispesifejä elementtejä, jotka vetoavat lähdekulttuurin vastaanottajiin, mutta jos ne käännetään sanatarkasti, ne menettävät merkityksensä. Esimerkiksi tutkielman aineistossa esiintyvät ilmaisut *kylmä Pohjolan talvi* ja *metsäseikkailut* eivät herätä samanlaisia tuntemuksia sellaisissa maissa, joissa metsä koetaan vaarallisena ja joissa talvet eivät ole kovin kylmiä. Tuotekuvausten, kuten muidenkin markkinointitekstien kääntämisen, voidaan tästä syystä nähdä poikkeavan käännöstieteessä usein vallitsevasta ekvivalenssiajattelusta, jossa kohdeteksti on uskollinen lähtötekstille. Markkinointiteksteissä uskollisuuden sijaan korostuu tekstin funktion säilyttäminen. (Torresi 2010: 23.) Tuotekuvausten kääntämisessä funktion säilyttämiseen pyritään mm. adaptaatiolla, jossa käännös sopeutetaan käyttökontekstiinsa niin, että kohdekulttuuriin tai kohdetekstin kontekstiin sopimattomat lähtötekstin piirteet muokataan sopivammiksi. Yllä mainittu esimerkki *kylmä Pohjolan talvi* on adaptoitu englanninkieliseen käännökseen *to cope with the cold of Scandinavian winters* sen funktion säilyttämiseksi.

Adaptaation lisäksi markkinointitekstien yhteydessä mainitaan usein luova kääntäminen tai sen englanninkielinen vastine *transcreation*. Luova kääntäminen on tyypillistä erityisesti markkinointitekstejä käännettäessä, sillä erilaiset markkinointi- ja mainoskampanjat vaativat usein tekstisisältönsä kokonaisvaltaista muokkaamista, jotta ne sopisivat kohdekulttuuriin. (Pedersen 2014: 58.)

Mainostekstien kieli ei ole täsmällistä eikä yksiselitteistä – kuten ei mikään muukaan luonnollinen kieli. Monitulkintaisuus on luonnollisen kielen ominaisuus, joka ilmenee esimerkiksi sanaston polysemiana: samalla sanalla on kaksi eri merkitystä tai useampia merkityksiä (Karlsson 2004: 213). Juuri monitulkintaisuus aiheuttaa kuitenkin ongelmia konekääntämisessä, jossa lyhyet, tiiviit ja yksiselitteiset lauseet tuottavat parhaimmat käännöstulokset (Mitamura 1999: 46). Luonnolliselle kielelle tyypillinen kontekstiriippuvainen tieto ja erot lähtö- ja kohdekulttuurin välillä vaikeuttavat konekääntämistä (Gross 1992: 109-110), samoin kuin pitkät ja monitulkintaiset lauseet (Calude 2003: 9). Tästä syystä luovien, etenkin kaunokirjallisten tekstien, on todettu aiheuttavan eniten haasteita konekääntimille. Toisaalta viimeaikaisessa tutkimuksessa on todettu, että neuroverkot suoriutuvat erityisen hyvin leksikaalisesti rikkaiden tekstien kääntämisestä (Bentivogli et al. 2016: 265).

Yllä kuvattua tekstien monitulkintaisuutta pyritään estämään erilaisin rajoituksin, jotka kohdistuvat tekstien syntaktisiin, semanttisiin sekä leksikaalisiin piirteisiin. Seuraavassa luvussa kuvataan tuotekuvauksissa esiintyviä rajoituksia ja pohditaan tuotekuvausten soveltuvuutta konekääntämiseen.

2.3.1 Rajoitukset tuotekuvauksissa

Jaaranen (2000) jakaa rajoitukset pro gradu -tutkielmassaan Lehrbergeriä (1982: 83) mukaillen kolmeen pääkategoriaan: leksikaalisiin, syntaktisiin ja semanttisiin rajoituksiin. Lehrberger tutki erikoiskielen kääntämistä englannista ranskaan, ja Jaaranen hyödynsi rajoitusten kategorisointia Elloksen tuotekuvausten rajoituksia kuvatakseen. Näin ollen myös tässä tutkielmassa rajoitukset jaetaan yllä mainittuihin kategorioihin. Rajoitusten esittelyllä pyritään muodostamaan yleiskuva tuotekuvauksiin liittyvistä rajoituksista, eli kyse ei ole sisällönanalyysista vaan pikemminkin yleistason havainnoinnista, jonka avulla tutkielman tuloksia pyritään tulkitsemaan myöhemmissä luvuissa.

2.3.2 Syntaktiset rajoitukset

Makrosyntagmat ovat kielen segmenttejä, joilla on yhtenäinen syntaktisten suhteiden muodostama sisäinen rakenne ja joiden välillä ei ole vastaavanlaisia syntaktisia suhteita (Loman & Jörgensen 1971: 9). Makrosyntagmat voivat olla muodoltaan interjektioita, puhutteluja, virkkeitä tai virkefragmentteja (Dannenberg 2004: 33). Tuotekuvausten makrosyntagmat ovat lähes aina virkefragmentteja, jotka koostuvat nominilausekkeista. Nominilausekkeet sisältävät etuattribuutteja, jotka ovat yleensä tuotetta kuvaavia adjektiiveja ja jälkimääritteitä (esim. adverbeja). Nominilausekkeen edussanana toimii useimmiten tuote, sen materiaali tai muu ominaisuus, kuten esimerkissä 3.

(3) a) Kaksi painonapeilla suljettavaa sivutaskua sekä koko etuosan halki kulkeva vetoketju.
b) An extremely warm and airy, lightweight insulation with natural down.

Tuotekuvauksissa on otettava huomioon tekstin pituutta koskevat rajoitukset. Vaikka tuotekuvausten pituudet vaihtelevat yrityskohtaisesti eikä merkkimäärille ole asetettu samanlaisia rajoituksia kuin esimerkiksi Twitterissä (280 merkkiä), on tuotekuvauksille varattu vain tietty tila tuotesivulla, sillä suurimman osan käytettävissä olevasta tilasta vie tuotekuva. Rajatun tilan takia nominilausekkeita ja etuattribuutteja yhdistelemällä pyritään ilmaisemaan tuotteen tärkeimmät ominaisuudet mahdollisimman tiiviissä ja yksinkertaisessa muodossa. Ymmärrettävyyden takaamiseksi tuotekuvaukset koostuvat usein yksinkertaisista lauseista, jotka sisältävät nominaalisen jäsenen (tai jäsenten) lisäksi finiittiverbin. Mainostekstilajeille tyypillisesti tuotekuvauksissa esiintyy kuitenkin myös syntagmoja, joista finiittiverbi puuttuu (esimerkit 3a ja 3b). Näitä lauseita kutsutaan vaillinaisiksi lauseiksi. Kirjoittajan ja lukijan yhteisen, yleisen ja tilannekohtaisen tiedon takia kaikkea ei tarvitse ilmaista, ja lauseet typistyvät. (Karlsson 2008: 122.) Vaillinaisten lauseiden käytöllä pyritään tuotekuvauksissa säästämään tilaa tuotteen kaikkein tärkeimpien ominaisuuksien kuvailuun.

Englanninkielisissä tuotekuvauksissa nominilausekkeiden lisäksi myös verbilausekkeet ovat yleisiä. Verbilausekkeiden funktiona on tuotteiden pääominaisuuksien esittämisen sijaan lukijaan vetoaminen ja lisätietojen antaminen:

c) The Quick Dry fabric can be washed in low temperatures, and it dries quickly.

Verbilauseke sopii kuitenkin parhaiten kuvaamaan kiinteän sanajärjestyksen omaavien kielten, kuten englannin, verbirakenteita. Sanajärjestyksen ollessa vapaampi kuten suomessa, muuttuu verbilausekkeen lausekestatus ongelmalliseksi. (Karlsson 2008: 138.)

2.3.3 Leksikaaliset rajoitukset

Tuotekuvaukset sisältävät tekstiilialan termejä, joita ei esiinny yleiskielessä, mutta jotka ovat tulkittavissa kontekstin ja esiintymisympäristönsä avulla. Lisäksi tuotekuvaukset poikkeavat muusta verkkosivuilla esiintyvästä tekstistä niin selkeästi, että lukija tunnistaa ne tuotekuvauksiksi.

Tuotekuvausten tuottamisessa ja kääntämisessä kaksi hyvin merkittävää osatekijää ovat termien oikeellisuus ja yhtenäisyys, joita Pasanen (2015: 120–121) kuvaa seuraavasti:

> ”Termien oikeellisuudella tarkoitan sitä, että käännöksessä puhutaan samoista käsitteistä kuin lähtötekstissä ja että näistä käsitteistä käytetään oikeita nimityksiä. Termien yhtenäisyydellä tarkoitan sitä, että käännöksessä käsitteestä käytetään systemaattisesti vain yhtä nimitystä, ellei synonyymisen nimityksenkäytölle ole perusteltuja syitä.”

Ongelmat termien yhtenäisyydessä ja oikeellisuudessa johtavat pahimmillaan kuluttajan harhaanjohtamiseen, kun esimerkiksi vettä hylkivää takkia markkinoidaan vedenpitävänä. Kuluttajansuojalain 6 §:n mukaan harhaanjohtavien tai totuudenvastaisten tietojen antaminen on kielletty markkinoinnissa, jos tiedot ovat omiaan johtamaan siihen, että kuluttaja tekee ostopäätöksen tai muun kulutushyödykkeeseen liittyvän päätöksen, jota hän ei ilman annettuja tietoja olisi tehnyt. Näin ollen virheet tuotekuvauksissa ja niiden käännöksissä saattavat johtaa reklamaatioihin ja lisäkuluihin, kun kuluttaja palauttaa vedenpitäväksi markkinoidun takin sen puutteellisten ominaisuuksien vuoksi.

Erikoisalan termistöön liittyvien rajoituksien lisäksi tuotekuvauksissa esiintyy myös muita leksikaalisia piirteitä, jotka toistuvat erilaisissa tuotekuvauksissa. Englannin kielessä määräisen artikkelin käytöllä pyritään erilaistamaan tuotetta tai sen valmistajaa ja korostamaan näin tuotteen erityisasemaa kilpailijoihin nähden (Torresi 2014: 128). Vastaavasti suomen kielessä demonstratiivipronomineilla voi luoda tekstiin lähentäviä (*tämä, nämä*) tai etäännyttäviä (*tuo, nuo*) sävyjä (Kielitoimiston ohjepankki: pronominit). Alla esimerkkejä aineistossa esiintyvästä määräisen artikkelin ja demonstratiivipronomien käytöstä.

(4) a) Create a trendy look with **these** Torstai cotton trousers.
b) You will feel stylish and warm sporting **this** ski jacket on the slopes.
c) To further enhance wearing comfort, **the** garment has been made with stretch material.
d) **Tämän** juoksupaidan Quick Dry -materiaali kuivuu nopeasti, minkä ansiosta mukava olo säilyy myös intensiivisten urheilusuoritusten aikana.

Yllä mainitut keinot keskittyvät tuotteeseen ja sen ominaisuuksien korostamiseen. Tuotteen sijaan pääpaino voi olla myös kuluttajassa, jolloin puhuttelussa käytetään tyypillisesti sinuttelua tai monikon ensimmäistä persoonaa, jotta mainostavan yrityksen ja kuluttajan välille syntyisi läheisempi yhteys. Puhuttelun lisäksi englannissa possessiivipronominia *(mine, yours*) ja suomessa persoonapronominien genetiivejä (*minun, sinun*) tai vaihtoehtoisesti substantiiviin liitettyä possessiivisuffiksia käytetään luonnollisen keskustelun vaikutelman luomiseksi. Myös imperatiivimuoto ja suorat tai retoriset kysymykset ovat hyvin tyypillisiä tuotekuvauksissa, sillä niiden käyttö vahvistaa yhteyttä mainostavan yrityksen ja kuluttajan välillä ja mahdollistaa lukijan suoran puhuttelun. (Torresi 2014: 128.)

Muiden mainostekstilajien tapaan myös tuotekuvauksissa esiintyvä kieli on usein luovien kielellisten tehokeinojen värittämää. Vapaamuotoisuudestaan huolimatta mainostekstilajit ovat kuitenkin hyvin tarkkaan säädeltyjä. Tiukasta sääntelystä vastaavat etenkin kuluttajansuojalaki ja sen tulkinnat. (Kuikka 2007: 39.)

2.3.4 Semanttiset rajoitukset

Tuotekuvauksia ei ole juurikaan tutkittu tekstilajina, mutta tutkielmassa voidaan olettaa, että eri sisällöntuottajien ja muiden tuotekuvauksia kirjoittavien henkilöiden vapaamuotoisissa tuotekuvauksissa on eroja. Luonteeltaan vapaata ja kertovaa tekstiä tarvitaan markkinoivia tekstejä kirjoittaessa, sillä tuotekuvausten tarkoituksena on tuoda lisäarvoa tuotteelle ja vakuuttaa kuluttaja tuotteen ominaisuuksista, jotta ostopäätös syntyisi. Vapaa teksti aiheuttaa kuitenkin ongelmia, sillä eroavaisuudet tuotekuvausten kirjoittajien käyttämässä termistössä ja tyylissä johtavat monitulkintaisuuteen ja vaikeuksiin vertailla eri tuotteita keskenään. Monitulkintaisuus ja epäyhtenevä termistön käyttö heikentävät myös konekääntimen tuloksia, sillä jos koulutusaineisto sisältää useita eri käännösvaihtoehtoja tietylle termille, pienenee todennäköisyys sille, että konekäännin päätyy valitsemaan oikean termin usean väärän joukosta.

Tuotekuvauksissa esiintyvien tekstiilialan termien on siis oltava yksiselitteisiä ja vakiintuneita. Teksteissä esiintyy kuitenkin usein myös yleiskielelle tyypillisiä sanoja. Niiden voidaan kuitenkin laskea kuuluviksi erikoiskieliseen sanastoon silloin, kun niillä on tuotekuvauksissa yleensä eri merkitys kuin yleiskielessä. Yleiskielisten sanojen, kuten *keskiosa*, *kerros* tai *olkapää*, merkityksen tulee siis olla pääteltävissä ympäröivästä kontekstista, jotta monitulkintaisuudelta vältytään. Ympäröivällä kontekstilla tarkoitan tuotekuvauksen lisäksi sen esiintymisympäristöä eli verkkosivustoa, joka sisältää tuotekuvauksen lisäksi myös valokuvia tuotteesta. Tuotekuvat ovat olennaisia, sillä ne tukevat tekstin tulkintaa.

(5) a) Olkapäät on vahvistettu synteettisellä, kosteutta torjuvalla toppauksella.
b) Reilunkokoinen malli, joka on helppo pukea muiden kerrosten päälle esimerkiksi tauoilla.
c) Keskiosassa vetoketjulliset, eristetyt taskut.

Yleiskielessä sana *olkapäät* viittaa vartalon osaan, mutta tässä yhteydessä kyseessä on takin osa. *Keskiosa* voi viitata mihin tahansa aina joen keskiosasta sohvan keskiosaan, mutta esimerkin 5c kontekstitiedon ja tuotekuvan perusteella kuluttaja voi päätellä, että kyseessä on takin keskiosa. Esimerkissä 5b yleiskielen sana *kerros* ei viittaa talon kerroksiin, vaan kontekstille tyypilliseen kerrospukeutumiseen. Konteksti sisältää siis yleensä tarpeeksi redundanssia oikean tulkinnan muodostumista varten (Karlsson 2008: 214). Erikoisalan termit erottuvat siis yleiskielestä juuri sopimustenvaraisien määritelmiensä vuoksi: termi merkitsee aina kontekstista riippumatta vakiintunutta asiaa, kun taas yleiskielisen sanan merkitys riippuu hyvin paljolti kontekstista (Pitkänen 2008: 122).

Erilaisten tekstien soveltuvuutta neuroverkkokääntämiseen on tutkittu toistaiseksi vain vähän, sillä neuroverkkokääntäminen on tutkimuksenalana vielä hyvin uusi. Tutkimuksen kohteena ovat toistaiseksi olleet esimerkiksi kielellisten piirteiden hyödyntäminen arvioidessa, miten erityyppiset tekstit soveltuvat neuroverkkokonekäännettäviksi (ks. Gröhn 2019) ja kaunokirjallisuuden kääntäminen neuroverkkokonekääntimellä (ks. Toral & Way 2018).

Vaikka erilaisten tekstilajien soveltuvuutta neuroverkkokääntimille on tutkittu vähän, voidaan silti todeta, että tuotekuvaukset sisältävät paljon rajoituksia, joiden ansiosta niiden konekääntämisen voidaan olettaa onnistuvan hyvin. Toisaalta tuotekuvaukset sisältävät myös niitä monimerkityksellisyyteen ja luovaan kielenkäyttöön liittyviä elementtejä, jotka aiheuttavat eniten ongelmia konekääntämisessä. Tuotekuvausten konekääntämisessä yhdistyvät siis

konekääntämisen heikkoudet ja vahvuudet – tästä syystä tuotekuvaukset ovatkin kiinnostava tutkimuskohde.

2.4 Neuroverkkokonekääntäminen

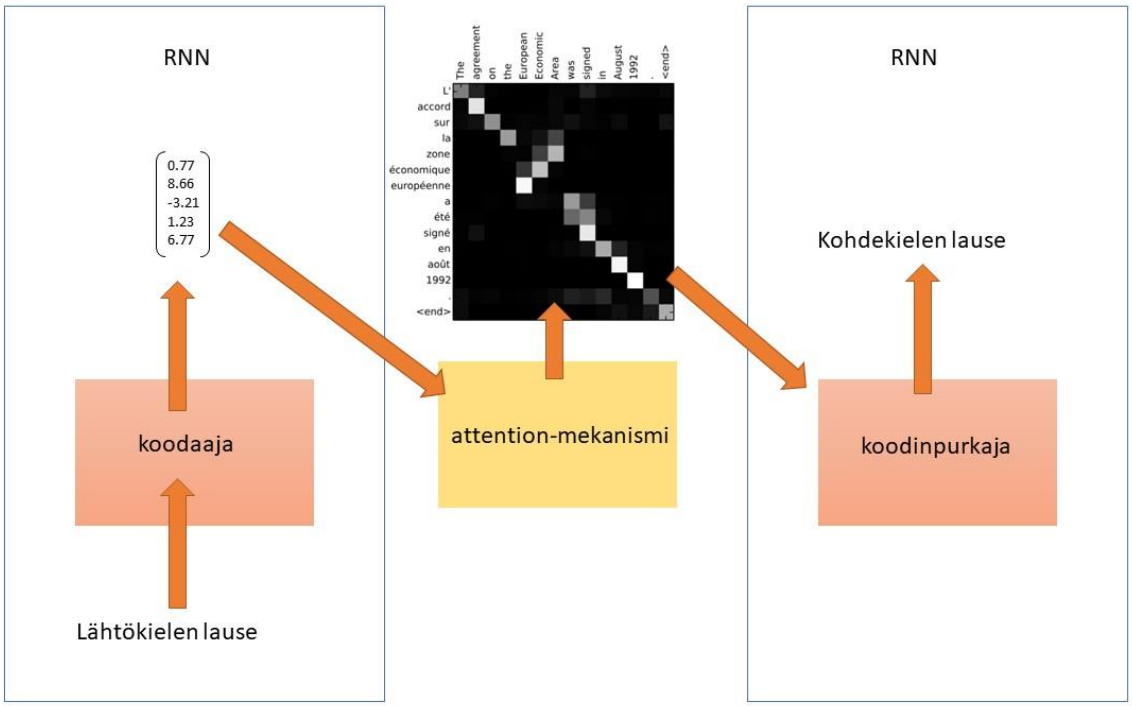
Käännösala on monen muun alan tavoin siirtymässä yhä teknisempään työskentelyotteeseen, jossa konekääntäminen – ja tällä hetkellä etenkin neuroverkkokonekääntäminen – ovat olennaisessa roolissa. Luonnollinen kieli on kuvien tunnistamisen ohella aiheuttanut eniten ongelmia tekoälylle, mutta neuroverkkoihin perustuvalla syväoppimisella molemmissa on saavutettu merkittäviä parannuksia. Neuroverkot ovat osoittautuneet tehokkaiksi luonnollisen kielen käsittelijöiksi, sillä ne voivat oppia monenlaisia tekstintuottamiseen, analysointiin ja kääntämiseen liittyviä tehtäviä. Neuroverkkoja on käytetty luonnollisen kielen käsittelyssä jo pitkään (Goldberg 2017: 3), mutta viimeisimmissä neuroverkkokonekäänninjärjestelmissä keskiössä ovat olleet erityisesti rekurrentit neuroverkot (Zhang & Zong 2015: 23) ja Vaswani et al:in (2017) esittelemä Transformer-arkkitehtuuri. Tässä luvussa esitellään ensin lyhyesti konekääntämisen kannalta merkittävää neuroverkkoteknologiaa, kuten koodaaja-koodinpurkaja-arkkitehtuuria ja attention-mekanismia, jonka jälkeen alaluvuissa käsitellään tarkemmin neuroverkkokonekääntämisen kannalta olennaiset verkkotyypit ja transformer-arkkitehtuuri.

Neuroverkkokonekääntäminen perustuu sekvenssimuotoiseen tiedonmallintamiseen: kun neuroverkkokonekääntimelle annetaan lähtökielinen sanasekvenssi, se tuottaa vastaavan sanasekvenssin kohdekielellä. Toiminta perustuu koodaaja-koodinpurkaja-arkkitehtuuriin: koodaaja lukee lähdekielen lauseen ja muuttaa sen vektorilistaksi, eli listaksi, jossa yksi vektori vastaa yhtä syötesymbolia. Tämän jälkeen koodinpurkaja käyttää vektorilistaa kohdekielisen lauseen tuottamiseen (ks. kuva 5). Suurin osa neuroverkkokääntimistä perustuu edellä kuvattuun koodaaja-koodinpurkaja-arkkitehtuuriin. (Sutskever et al. 2014; Cho et al. 2014a.)

Sekä koodaaja- että koodinpurkajaverkot ovat aikaisemmin olleet rekurrentteja neuroverkkoja (Cho et al. 2014a), eli ne hyödyntävät rekursiivista oppimista, jossa seuraava arvioitu lopputulos riippuu edellisistä arvoista. Tämä muistiominaisuus on erityisen hyödyllinen luonnollisen kielen käsittelyssä. Koodaaja-koodinpurkaja-arkkitehtuurin täydennykseksi kehitettiin hyvin nopeasti attention-mekanismi (Bahdanau, Cho & Bengio 2016), joka mahdollistaa sen, että koodinpurkaja

voi yksittäisiä sanoja kääntäessään kiinnittää huomiota lähdekielisen lauseen eri osiin (Forcada 2017: 8), parantaen näin etenkin pitkien lauseiden käännöslaatua. Tällä hetkellä suosituin arkkitehtuuri on kuitenkin eteenpäin syöttäviin neuroverkkoihin perustuva Transformer-arkkitehtuuri, joka poisti tarpeen rekursiiviselle oppimiselle monipäisellä tarkkaavaisuusmekanismillaan. Transformer-arkkitehtuuria esitellään tarkemmin luvussa 2.4.2.

Attention-mekanismi on koodaaja-koodinpurkaja-arkkitehtuurin laajennos, jonka toiminta perustuu siihen, että sanoja linjataan (eng. *align*) ja käännetään samanaikaisesti. Mekanismissa koodinpurkaja muodostaa useita vektoreita lauseen eri sanoille. Aina kohdekielistä sanaa tuotettaessa attention-mekanismi etsii ne kohdat lähtökielisestä lauseesta, joihin olennaisin tieto on keskitetty. Niitä kontekstivektoreita, jotka sisältävät olennaisimman tiedon, käytetään kohdekielisen sanan ennustamiseen. Lisäksi kohdekielisen sanan ennustamiseen käytetään jo siihen mennessä käännettyjä sanoja. (Bahdanau, Cho & Bengio 2016: 1.)



Kuva 5. Yksinkertaistettu kuvaus koodaaja-koodinpurkaja-arkkitehtuurista attention-mekanismilla (Bahdanau, Cho & Bengio 2016: 6).

2.4.1 Rekurrentit neuroverkot ja LSTM-verkot

Luonnollisen kielen ongelmissa, kuten kääntämisessä, syötteet pilkotaan sekvensseiksi (esim. sanoiksi tai kirjaimiksi), sillä syötteet ovat usein liian suuria ja monimutkaisia tulkittaviksi sellaisinaan. Sekvenssit sisältävät paljon kontekstiriippuvaista tietoa, jonka avulla esimerkiksi tiettyjen kirjainten sarja voi muodostaa sanan. Juuri tällaista sekvenssimuotoisen tiedon prosessointia varten on kehitetty rekurrentteja neuroverkkoja (eng. *Recurrent Neural Network, RNN*). Rekurrentteja neuroverkkoja käytetään jaksomaisten kaavojen tai tapahtumien ennustamiseen: käsiteltävä data sisältää usein tietoa, joka on toistuvaa tai edellisestä asiasta riippuvaista. Rekurrentit neuroverkot pystyvät siis muistamaan aiemmin tapahtuneita asioita. RNN voidaan opettaa esimerkiksi tulkitsemaan kirjainsekvenssiä siten, että se pystyy ennustamaan seuraavaksi tulevan kirjaimen kahden edellisen kirjaimen perusteella. (Helenius 2016: 10.) Kääntämisen ohella rekurrentteja neuroverkkoja voidaan käyttää mm. puheentunnistamiseen tai osakemarkkinoiden ennustamiseen.

Kaikkien neuroverkkotyyppien tapaan myös rekurrentissa neuroverkossa on syötekerros, piilotetut kerrokset sekä tulostekerros. Kerrosten lisäksi rekurrenteissa neuroverkoissa piilotettujen kerroksien neuroneilla on silmukoita sekä yhteyksiä takaisin itseensä (Jaokar et al. 2015). Piilokerros ottaa aiemman piilokerroksen arvon ja tämänhetkisen syötteen, ja laittaa ne aktivaatiofunktion läpi. RNN:n rakenteen ja yksittäisissä neuroneissa sijaitsevien muistisolujen ansiosta edelliset syöttötiedot voidaan säilyttää ja niitä voidaan hyödyntää lopputuloksen tuottamiseen missä tahansa vaiheessa. Näin samanlaista tasoa käytetään jokaisessa vaiheessa, ja edellisen vaiheen lopputulosta käytetään seuraavan vaiheen syöttötietona. (Graves 2012: 18). Seuraava solu tietää siis edellisen solun tulosteen, jota se hyödyntää sen hetkistä tulostetta luodessaan. Näin tulosteen luomisessa otetaan aina huomioon myös aiemmin huomioidut asiat.

Joskus tulosteen luomiseen tarvitaan vain viimeisintä, äskettäin esiintynyttä tietoa. Kun kielimalli yrittää ennustaa seuraavaksi tulevaa sanaa edellisen perusteella, on lause ”*the clouds are in the **sky**”* sille suhteellisen helppo, sillä sanan ”*sky*” ennustamiseen ei tarvita laajempaa kontekstia. Tällaisissa tapauksissa tarvittava aikaisempi tieto ja sen käyttöhetki ovat lähekkäin, ja RNN oppii hyödyntämään aikaisempaa tietoa tehokkaasti. Jos seuraavan sanan ennustamiseen tarvitaan kuitenkin enemmän kontekstia, kuten lauseessa ”*I grew up in France… I speak fluent **French***”,

hankaloituu ennustaminen. Viimeisimmän syötetiedon mukaan seuraava sana on luultavasti jokin kieli, mutta jos mahdollisten kielten määrää halutaan rajata, tarvitaan ennustamiseen jo aikaisemmin esiintynyt tieto Ranskasta (*France*). Näin ollen tarvittava tieto ja sen käyttöhetki saattavat sijaita kaukana toisistaan. Mitä kauempana tarvittava tieto sijaitsee, sitä huonommin RNN oppii yhdistämään tietoa ja konteksti katoaa. (Olah 2015.).

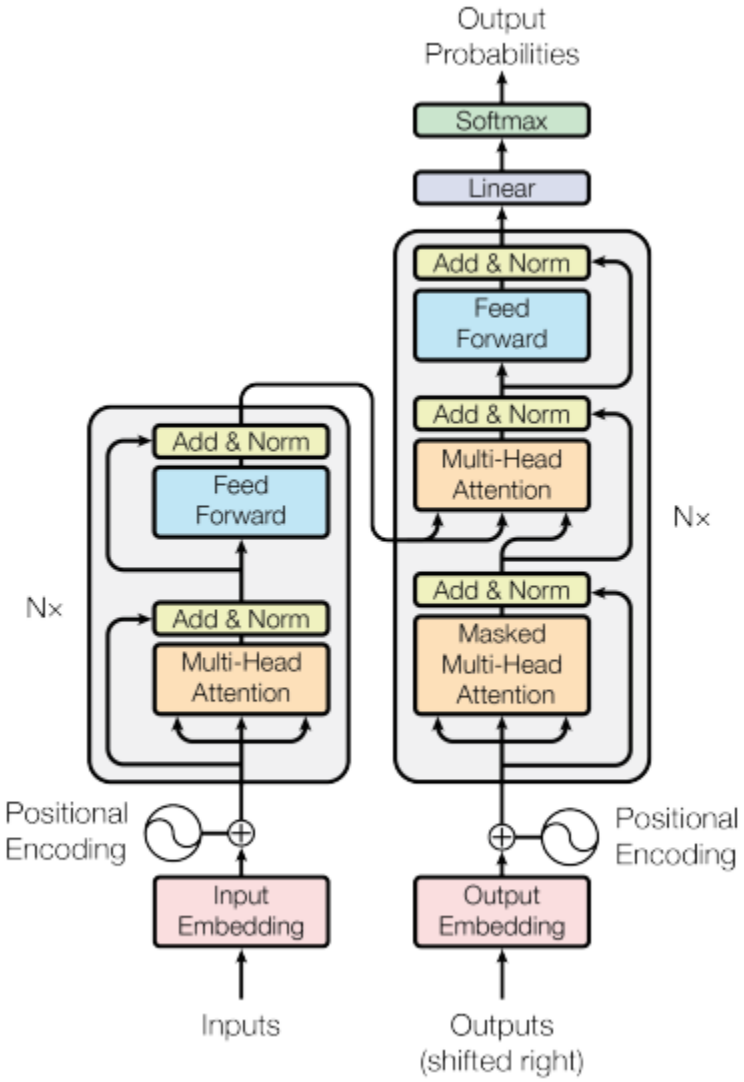
Vaikka rekurrentit neuroverkot on tehty luokittelemaan kontekstiriippuvaista tietoa, ne saattavat kadottaa kontekstin etenkin pidemmillä sekvensseillä, kun verkossa on paljon piilokerrosten askelia. RNN-verkko pystyy siis muistamaan vain lyhytaikaisia tapahtumia ja riippuvuuksia. Tämä johtuu katoavan gradientin ongelmasta (engl. *vanishing gradient*), joka tulee vastaan rekurrentin neuroverkon opetusprosessissa (Manaswi 2018: 117-118, lisätietoa opetusprosessista ks. esim. Lipton, Berkowitz & Elkan 2015).

Toistaiseksi tehokkain ratkaisu rekurrenttien neuroverkkojen pitkän aikavälin oppimisen puutteisiin on ollut LSTM-malli (eng. *Long short term memory, LSTM*) (Hochreiter & Schmidhuber 1995: 6–7). LSTM on rekurrentin neuroverkon malli, johon on lisätty muistielementtejä eli LSTM-soluja, jotta se pystyy oppimaan riippuvuuksia pitkältäkin aikaväliltä. (Geron 2017: 400–402.) Konekääntämisessä tämä on tärkeää, sillä aiempi tieto vaikuttaa aina tulevaan tietoon. Konekääntämisen lisäksi LSTM-mallia hyödynnetään myös esimerkiksi puheen- ja käsialantunnistuksessa.

LSTM-verkon piilokerros on rekurrenttien neuroverkkojen piilokerroksia monimutkaisempi, sillä muistisolu on sisällytetty kuhunkin piilokerroksen neuroniin. LSTM-solusta on kehitetty useita variaatioita, mutta yleisimmin käytetyssä LSTM-versiossa solu koostuu kolmesta portista, jotka ovat unohtamisportti (*forget gate*), syöteportti (*input gate*) ja tulosportti (*output gate*). (Greff, Srivastava, Koutník, Steunebrink & Schmidhuber 2015: 1.) LSTM-verkon piilokerroksessa ylläpidetään erityistä solun tilaa (*cell state*), joka erikoistuu pidemmän etäisyyden riippuvaisuuksiin sekvenssissä. Tätä muokataan sekvenssin elementtejä syöttäessä käyttämällä edellä mainittuja portteja, joita käyttämällä pyritään säilyttämään solun tilassa tärkeät riippuvaisuussuhteet. (Helenius 2016: 10.) LSTM-solu lisää siis signaalin kulkuun portteja ja tiloja, jotka päättävät mitä tietoja aiemmista soluista unohdetaan tai muistetaan, mitä tietoja lisätään sen hetkiseen syötteeseen sekä mitä tietoja annetaan vasteena (Lipton, Berkowitz & Elkan 2015: 17).

2.4.2 Transformer-neuroverkkoarkkitehtuuri

Neuroverkot voidaan luokitella kahteen eri tyyppiin niiden verkkotopologian perusteella: edellä käsiteltyihin rekurrentteihin eli takaisinkytkettyihin neuroverkkoihin ja eteenpäin syöttäviin neuroverkkoihin (eng. *feed-forward network, FFN*). Transformerissa hyödynnetään rekurrentin neuroverkon sijaan eteenpäin syöttävää neuroverkkoa, jossa data kulkee vain yhteen suuntaan (Vaswani et al. 2017: 2). Rekursiiviselle oppimiselle ei ole enää tarvetta, kun eteenpäin syöttävään neuroverkkoon yhdistetään monipäinen huomiomekanismi.



Kuva 6. Transformerin arkkitehtuuri (Vaswani et al. 2017: 3).

Transformerin arkkitehtuuri koostuu kuudesta identtisestä koodaaja-koodinpurkajakerroksesta (Vaswani et al. 2017: 2). Arkkitehtuurissa hyödynnetään monipäistä huomiomekanismia (eng. *Multi-Head Attention*), joka suorittaa tarkkaavaisuusfunktion usealle eri kyselylle yhtä aikaa. Näitä monipäisiä huomiomekanismeja on Transformerissa kolme, ja ne on esitetty kuvassa 6 oransseilla laatikoilla. Yksi on koodaajassa ja kaksi muuta sijaitsevat koodinpurkajassa. Yksi koodinpurkajassa sijaitsevista monipäisistä huomiomekanismeista on yhteydessä koodaajaan, jotta koodaajan syöte saadaan siirrettyä koodinpurkajaan. Tämä mahdollistaa sen, että koodinpurkaja pystyy kiinnittämään lauseessa huomiota kohtiin, jotka tarjoavat eniten kontekstia kulloinkin käsitellylle sanalle (Vaswani et al. 2017: 3–5) eikä rekursiolle ole enää tarvetta.

Bahdanau et al.:in (2016) esittelemän attention-mekanismin jälkeen Vaswani et al. (2017) esittelivät self-attentioniin perustuvan huomiomekanismin: Transformerin monipäisessä huomiomekanismissa hyödynnetään self-attention-kerroksia, joita on sijoitettu sekä koodaajaan että koodinpurkajaan. Self-attentionin erikoisuus on se, että se jättää huomioimatta sanojen välisen etäisyyden ja laskee suoraan riippuvuussuhteet, jolloin se pystyy oppimaan lauseen sisäisen rakenteen. Self-attentioniin perustuvat huomiomekanismit ovat tällä hetkellä suosittuja tutkimuskohteita ja niiden käyttöä tutkitaan erilaisissa luonnollisen kielen käsittelytehtävissä (ks. esim. Im & Cho 2017). (Cloud 2018.)

Arkkitehtuurinsa, etenkin monipäisen huomiomekanisminsa, ansiosta Transformerin koulutus on huomattavasti nopeampaa kuin rekurrentteihin- tai konvoluutiokerroksiin perustuvien arkkitehtuurien koulutus (Vaswani et al. 2017: 9). Tästä syystä Transformer on tällä hetkellä hyvin suosittu arkkitehtuuri konekääntämisessä.

Vaikka neuroverkkoteknologia on mullistanut konekääntämisen ja sen kehityksen, liittyy neuroverkkokonekääntämiseen silti vielä paljon ongelmia, joiden ratkaisemiseksi tarvitaan lisää tutkimusta. Pitkien ja monimutkaisten lauseiden lisäksi esimerkiksi pienet koulutussanastot aiheuttavat ongelmia neuroverkkokonekäänninjärjestelmille (Cho, van Merriënboer & Bahdanau 2014b: 3–4). Myös laatu aiheuttaa ongelmia, sillä erityisesti kielipareissa, joissa laajoja rinnakkaiskorpuksia ei ole saatavilla, vaikeutuu myös neuroverkon automaattinen oppiminen (Shen et al. 2015: 1–2).

3 Aineisto ja tutkimusmetodi

Tässä luvussa kuvataan aineiston keruuta ja käsittelyä, jonka jälkeen esitellään tutkielmassa koulutettu neuroverkkokonekäännin Marian ja konekäännösten automaattiseen evaluointiin käytetty BLEU-menetelmä. Itse aineistoa on esitelty tarkemmin luvuissa 2.2 ja 2.3.

3.1 Aineiston keruu ja käsittely

Markkinointiviestinnän muuttuessa yhä visuaalisemmaksi verkkosivustoilta löytää erilaisten tekstien lisäksi muutakin sisältöä, kuten esimerkiksi tuote- ja brändikuvia, grafiikkaa, videoita ja taulukoita (Kim & Kuljis 2010: 369–370). Tässä tutkimuksessa käytetty aineisto koostuu ainoastaan tuotekuvausten ja ominaisuusluetteloiden tekstisisällöstä, ja aineiston ulkopuolelle on jätetty esimerkiksi tuotekuvaustekstien yhteydessä esiintyvät tuotteiden kuvat ja tuotenimet sekä mahdolliset muut erillisten linkkien takaa löytyvät lisätiedot.

Toimiva viestintä edellyttää verkkosivujen jatkuvaa ylläpitoa (Isohookana 2007: 275). Verkkosivuaineistoja tutkittaessa onkin otettava huomioon, että verkkosivujen sisältöjä päivitetään usein. Tekstiilialan yritykset päivittävät verkkokauppojensa tuotteet sesongeittain, mutta verkkosivujen muuta sisältöä lisätään ja muokataan viikoittain. Aineisto tulee tästä syystä kerätä lyhyen aikavälin sisällä, jotta verkkosivustojen sisällöt eivät ehdi muuttua kesken aineistonkeruuprosessin (Kim & Kuljis 2010: 370). Tämän tutkielman aineisto kerättiin yhdeksän päivän aikana marraskuussa 2019. Aineiston keruun ajankohdan vuoksi aineisto koostuu syys- ja talvivaatteiden tuotekuvauksista eikä näin ollen sisällä kevät- tai kesäsesongin tuotteita.

Kaksikielisen koulutusmateriaalin saatavuus on yksi suurimmista ongelmista, joita neuroverkkojärjestelmien kouluttamiseen ja kehittämiseen liittyy. Ongelmaan on esitetty muutamia lupaavia ratkaisuja, kuten synteettisen datan käyttö, jossa olemassa olevasta yksikielisestä materiaalista valitaan uudelleenkoulutuksen kannalta sopivimmat segmentit, jotka konekäännetään kaksikielisen korpuksen luomiseksi (ks. esim. Chinea-Ríos, Peris & Casacuberta 2017: 138–145). Tässä tutkielmassa neuroverkkokääntimen uudelleenkoulutukseen ei käytetä olemassa olevaa korpusta tai synteettistä dataa, vaan aineisto koostetaan vapaasti internetistä saatavilla olevista kaksikielisistä tuotekuvauksista, jotka on kerätty html-muotoisina wget-apuohjelmalla ja linjattu. Näin varmistetaan, että uudelleenkoulutusmateriaali ei sisällä samoja

tekstejä, kuin esikoulutetun konekääntimen koulutusmateriaali. Lisäksi tarkoituksena oli tutkia, kuinka helposti yksityishenkilö, esimerkiksi freelancer-kääntäjä, saisi itse koostettua oman kaksikielisen korpuksensa nettisivujen materiaalista. Aineiston keräyksen tuloksena saatua kaksikielistä korpusta voidaan hyödyntää myös muuhun tutkimustarkoitukseen, kuten tässä tapauksessa esimerkiksi terminologiseen tutkimukseen.

Aineisto sisältää englannin- ja suomenkielisiä tuotekuvauksia seuraavilta brändeiltä: Luhta, Icepeak, Rukka, Torstai, Marimekko, Minna Parikka, Fjällraven, Haglöfs, Lovia, Didriksons, Joutsen ja Revolution Race. Neljällä ensimmäisellä brändillä tuotekuvaukset ovat lähtökieleltään suomenkielisiä, mutta muiden brändien osalta lähtökielestä ei ole varmuutta. Aineistoa kerättäessä huomiota kiinnitettiin erityisesti laatuun ja siihen, että kyseessä on nimenomaisesti käännös eikä kieliversio. Koska kyseessä ovat usean eri brändin käännökset, sisältää kerätty materiaali varmasti joitain käännösvirheitä ja vaihtelua käytetyssä terminologiassa, kuten luvussa 2.3.4 todettiin.

Aineiston keräyksen jälkeen englannin- ja suomenkieliset tiedostot linjattiin käännösmuistiohjelma Memsourcen linjaustyökalulla, joka mahdollistaa usean lähtö- ja kohdekielisen tiedoston linjaamisen samanaikaisesti. Tämän jälkeen linjatusta tiedostosta poistettiin manuaalisesti kaikki tuotekuvauksiin liittymätön sisältö ja linjausten paikkansapitävyys tarkastettiin.

Linjaamisen ja siistimisen jälkeen n. 5000 segmentin aineisto jaettiin kahdeksi erilliseksi tiedostoksi (lähtö- ja kohdekieli eli tässä tutkielmassa suomi-englanti). Suomenkielistä materiaalia käytettiin lähtökielenä ja englanninkielistä materiaalia käytettiin referenssimateriaalina BLEU-pisteiden arvioinnissa.

Ennen konekääntämistä suomenkielinen teksti palasteltiin (eng. *Tokenisation*) Sentencepiece-mallilla, jossa teksti palastellaan osiksi, jotka eivät ole mitään tiettyjä kokonaisuuksia, kuten sanoja tai tavuja. Sentencepiece-malli perustuu valvomattomaan oppimiseen, jossa koneelle syötetään vain luokittelematonta dataa, jota algoritmit käyvät läpi itsenäisesti. Palastelemalla konekääntimen tarvitsema sanasto saadaan rajattua noin 30.000 yleisimpään merkkijaksoon (eng. *Token*), joista kaikki aineistossa esiintyvät sanat voidaan koostaa. Sentencepiece on kehitetty lähinnä neuroverkkopohjaisia tekstinmuodostusjärjestelmiä varten ja sitä voidaan käyttää missä tahansa kielessä edellyttäen, että se koulutetaan kieliparikohtaisesti. (Domingo, García-Martínez, Helle, Casacuberta & Herranz 2018: 3.)

Kun suomenkielinen lähtöteksti oli palasteltu ja konekäännetty, laskettiin englanninkieliselle konekäännökselle BLEU-arvo (ks. luku 3.4.1), jotta tulosta voitaisiin myöhemmin vertailla uudelleenkoulutetun konekääntimen saamaan BLEU-arvoon.

3.2 Konekääntimen koulutus kerätyllä korpuksella

Neuroverkkokääntimen kouluttaminen vaatii todella suuren määrän koulutusdataa, mutta se myös oppii datasta huomattavasti edeltäjiään tehokkaammin. NMT-järjestelmät ovat myös yleensä hyviä adaptoitumaan tietyn aihealueen teksteihin, mutta laatu kärsii heti, kun kääntimellä käännetään aihealueeseen kuulumattomia tekstejä. Kun konekäännin koulutetaan geneerisellä aineistolla, eli aineistolla, joka sisältää kaikenlaisia tekstejä aina kaunokirjallisuudesta tekstityksiin ja uutisartikkeleihin, suoriutuu se heikommin teksteissä, jotka käsittelevät esimerkiksi tiettyä erikoisalaa. (Koehn & Knowles 2017: 1–2.) Ratkaisuna tähän konekääntimiä on alettu kouluttamaan geneerisellä koulutusdatalla ja vasta sen jälkeen mukaan koulutukseen on otettu tietyn aihealueen tekstejä, joiden avulla konekäännin saadaan suoriutumaan paremmin ja ns. erikoistumaan tiettyyn erikoisalaan, vaikka se pohjakoulutukseltaan onkin geneerinen konekäännin (ks. esim. Freitag ja Al-Onaizan 2016, Luong & Manning 2015 ja Mäkinen 2019). Samaa metodia käytettiin myös tässä tutkielmassa.

Esikoulutettua neuroverkkoa uudelleen koulutettaessa on otettava huomioon, että saadut tulokset pohjautuvat paljolti siihen dataan, jolla verkko on esikoulutettu. Tutkielman laajuuden ja laskentaresurssien vuoksi valmiiseen malliin tukeutuminen oli kuitenkin kannattavampaa kuin verkon kouluttaminen alusta asti, sillä tutkielmassa tarkastellaan nimenomaisesti uudelleenkoulutuksen tehokkuutta, kun kyseessä on hyvin rajattu aineisto. Samaan metodia ovat käyttäneet mm. Freitag ja Al-Onaizan (2016), jotka uudelleenkouluttivat esikoulutettua neuroverkkoa pienellä aihealuespesifillä koulutusaineistolla parantaakseen sen suoriutumista tietyn aihealueen käännöksissä. He yhdistivät esikoulutetun ja uudelleenkoulutetun mallin, jotta ylikoulutukselta ja muiden kuin aihealuespesifien segmenttien laadun heikentymiseltä vältyttäisiin. Ylikouluttamisella tarkoitetaan tilannetta, jossa konekäännin suoriutuu erinomaisesti koulutusaineiston kaltaisen materiaalin kääntämisestä, mutta suoriutuminen heikkenee merkittävästi muunlaista materiaalia käännettäessä. Freitag ja Al-Onaizan huomasivat, että BLEU-pisteiden nousuun ei vaadittu useita koulutuskertoja: käännöslaatu parani saksa-englanti

kieliparissa kahden koulutuskerran jälkeen ja kiina-englanti kieliparissa kuuden koulutuskerran jälkeen. (Freitag & Al-Onaizan 2016: 3–8.)

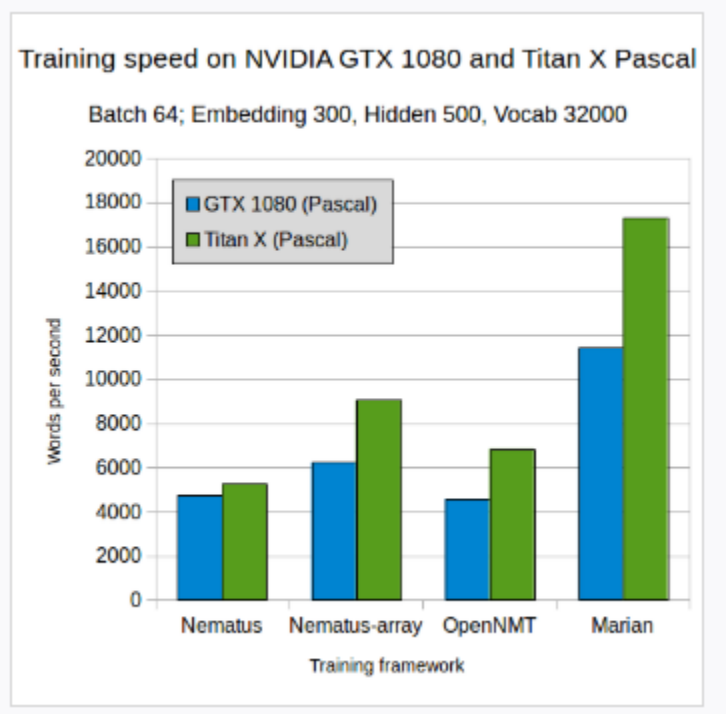
Kun aineisto oli jaettu lähtö- ja tulokielisiin tiedostoihin, jaettiin tiedostot vielä tämän jälkeen erillisiksi koulutus- ja testaustiedostoiksi k-ositetulla ristiinvalidoinnilla (eng. *k-fold cross-validation*), jossa aineisto jaetaan k:hon yhtä suureen osaan. Tässä tutkielmassa aineisto jaettiin kymmeneen yhtä suureen osaan (k = 10). Jako tehtiin, jotta testausaineistoa ei käytettäisi kouluttamiseen, sillä koulutusaineiston käyttö testauksessa nostaisi BLEU-pisteitä huomattavasti.

Jaon jälkeen professori Jörg Tiedemannin Opus-korpuksen aineistolla (ks. Tiedemann 2012) esikouluttama neuroverkkokonekäännin Marian (ks. luku 3.3) jatkokoulutettiin 90%:lla tuotekuvauksista koostuvasta aineistosta, jonka jälkeen testiaineisto (loput 10% aineistosta) käännettiin koulutetulla mallilla. Tämän jälkeen käännöksille laskettiin BLEU-pisteet, joita verrattiin esikoulutetulla mallilla saatuihin tuloksiin.

3.3 Tutkimuksessa käytetty neuroverkkokäännin

Kuten aikaisemmissa luvuissa todettiin, sanoilla on useita eri käännösvaihtoehtoja etenkin, kun liikutaan eri aihealueiden ja erikoiskielten välillä. Tästä syystä on tärkeää, että tiettyyn tarkoitukseen kehitettävät konekääntimet koulutetaan kyseessä olevalle erikoisalalle. Tällä hetkellä neuroverkkokääntimiä koulutetaan erikoisalalle niin, että geneerisellä materiaalilla esikoulutettu konekäännin jatkokoulutetaan erikoisalan materiaalilla. (Koehn & Knowles 2017: 2, ks. myös Luong and Manning 2015; Freitag & Al-Onaizan 2016.)

Tässä tutkielmassa käytettiin Helsingin yliopiston kieliteknologian professorin Jörg Tiedemannin esikouluttamaa Marian-neuroverkkokonekäännintä. Esikoulutusmateriaalina on käytetty Opus-korpusta (ks. Tiedemann 2012). Marian on parhaillaan käytössä useissa eurooppalaisissa projekteissa ja se toimii mm. maailman henkisen omaisuuden järjestön neuroverkkokonekääntämishankkeen pääasiallisena kääntämis- ja koulutusjärjestelmänä (Junczys-Dowmunt et al. 2018: 116). Lisäksi Mariania hyödynnetään useassa tutkimusprojektissa (esim. FISKMÖ) ja useissa yrityksissä, kuten Microsoft Translator -neuroverkkokäännöspalvelun konekäänninjärjestelmänä. (MarianNMT, 2018).



Kuva 7. Marianin koulutusnopeus verrattuna muihin järjestelmiin. (MarianNMT, 2018).

Marian NMT perustuu koodaaja-koodinpurkaja-arkkitehtuuriin, jossa hyödynnetään luvussa 2.4 kuvattua attention-mekanismia ja Transformer-arkkitehtuuria. (Junczys-Dowmunt et al. 2018: 117). Marian NMT:n arkkitehtuuri vastaa läheisesti Nematuksen arkkitehtuuria (ks. Senrich et al. 2017), sillä Marian syntyi Nematuksen C++-uudelleentoteutuksesta (Junczys-Dowmunt et al. 2018: 116). Kuten kuvasta 8 on havaittavissa, Marianin koulutus on nopeampaa kuin sen edeltäjien, sillä esimerkiksi toisin kuin OpenNMT:ssä, Marianissa koulutus on mahdollista suorittaa usean grafiikkaprosessorin (eng. *graphics processing unit*, *GPU*) avulla. Marian on noin neljä kertaa nopeampi yhdellä GPU:lla koulutettaessa ja 30 kertaa nopeampi 8 GPU:lla koulutettaessa identtisiin malleihin verrattaessa (MarianNMT, 2018). Muun muassa tästä syystä tutkielma toteutettiin OpenNMT:n sijaan Marianilla.

3.4 Konekäännösten automaattinen arviointi

Konekäännösten laadun arviointiin on kehitetty erilaisia menetelmiä 1950-luvulta lähtien konekääntämisen yleistyttyä. Laadunarviointimenetelmät olivat aluksi lähes poikkeuksetta manuaalisia, eli ne perustuivat kohdekielen osaajien arvioihin konekäännöksen laadusta. Korkeiden kustannusten lisäksi manuaalinen arviointi on hidasta: ihmisten tekemää arviointiin saattaa kulua viikkoja ja tai kuukausia, vaikka laadunarviointi on konekäännösjärjestelmien kehityksen ydin ja hitaat arviointimenetelmät johtavat usein pullonkaulaan konekäännöstutkimuksen edistymisessä. Automaattinen evaluointi onkin syntynyt vastaamaan tutkijoiden tarpeeseen edullisesta laadunarvioinnista, joka on kieliparista riippumatonta, nopeaa ja vastaa ihmisten tekemää arviointia. Tutkijat hyötyvät edullisesta laadunarvioinnista, joka on nopeaa, kieliparista riippumatonta ja vastaa läheisesti ihmisten suorittamaa arviointia. (Papineni et al. 2002: 311). Seuraavassa luvussa esitellään tutkielmassa käytettyä automaattisen laadunarvioinnin menetelmää.

3.4.1 BLEU

Tässä tutkielmassa konekäännösten laadunarviointiin käytetään BLEU-menetelmää. BLEU on yksi eniten konekäännösten laadunarviointiin käytetyistä mittareista, ja sen tulokset vastaavat läheisesti ihmisten tekemiä arviointeja (Lin & Och 2004: 1). Tässä kappaleessa esitellään BLEU:n toimintaperiaatteita ja rajoituksia.

BLEU:ssa konekäännöksen n-grammeja verrataan referenssikäännösten n-grammeihin ngram-rinnakkaistilastojen avulla, jonka jälkeen BLEU laskee geometrisen keskiarvon n-grammien tarkkuudesta. BLEU antaa yksittäisille segmenteille pisteitä asteikolla 0–1 segmenttien laadun perusteella, ja laskee sitten kaikkien segmenttien keskiarvon koko käännöksen arvioksi. Arvioinnin lähtökohtana on, että mitä lähempänä konekäännös on ihmisen tekemää referenssikäännöstä, sen parempi konekäännös on. Jos arvioitavana olevalla käännöksellä ei ole yhtään n-gram osumaa, on tällä käännöksellä virkettä kohden 0 BLEU-pistettä. Tuloksena saatu BLEU-piste ilmaisee siis sitä, kuinka hyvin konekäännös vastaa referenssikäännöstä. (Papineni et al. 2002: 311–315.)

BLEU mittaa siis käännöksen tarkkuutta (eng. *precicion*) vertaamalla konekäännöstä yhteen tai useampaan ihmisen tekemään referenssikäännökseen. Täsmällisyys määritetään laskemalla konekäännöksen ne sanat (unigrammit), jotka esiintyvät myös missä tahansa referenssikäännöksessä. Tämä jälkeen sanamäärä jaetaan konekäännöksen kokonaissanamäärällä. On kuitenkin havaittu, että konekäännökset sisältävät usein enemmän sanoja kuin referenssikäännökset, sillä konekääntimet ylituottavat ”järkeviä” sanoja, jotka johtavat täsmällisyysarvoltaan korkeisiin, mutta käyttökelvottomiin käännöksiin. (Papineni et al. 2002: 312.) Tästä syystä BLEU:ssa käytetään muokattua n-grammin tarkkuusmittaa, jossa konekäännökset, jotka ovat referenssikäännöksiä pidempiä, saavat alhaisemman BLEU-pisteen. N-grammin tarkkuus lasketaan seuraavasti:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)}$$

Kaava 1. N-grammin täsmällisyysmitta. (Papineni et al. 2002: 313).

Kaaviossa $Count_{clip}$ (n-grammi) on konekäännöksessä ja referenssikäännöksessä esiintyvien n-grammien maksimimäärä ja *Count* (n-grammi) on konekäännöksessä esiintyvien n-grammien määrä.

Jotta BLEU ei suosisi lyhyitä käännöksiä, algoritmiin on lisätty myös lyhyyssakko (eng. *Brevity Penalty, BP*), jonka mukaan korkeiden BLEU-pisteiden saaminen edellyttää, että konekäännöksen on vastattava referenssikäännöstä pituudeltaan, sanavalinnoiltaan ja sanajärjestykseltään. Muokattu n-grammin tarkkuusmitta tai lyhyyssakko eivät kuitenkaan huomioi lähtötekstin pituutta vaan ne ottavat huomioon vain kohdekielisten referenssikäännösten pituudet. (Papineni et al. 2002: 315.) Lyhyyssakko (BP) lasketaan seuraavasti:

$$BP = \begin{cases} 1 & if \ |c| > |r| \\ e^{(1-|r|/|c|)} & if \ |c| \leq |r| \end{cases}$$

Kaava 2. Brevity Penaltyn laskuyhtälö (Papineni et al. 2002: 315).

Kaaviossa |c| on konekäännöksen pituus ja |r| on referenssikäännöksen pituus. Lyhyyssakkoa ei kuitenkaan lasketa lausekohtaisesti niin, että yksittäisille lauseille laskettaisiin arvo ja lopuksi kaikista arvoista laskettaisiin keskiarvo, sillä lyhyiden lauseiden pituuspoikkeamat laskisivat BLEU-pisteitä. Lyhyyssakko lasketaan siis koko korpukselle, jotta lausetason vapaus säilyisi. (Papineni et al. 2002: 315.) Korpustason BLEU-pisteiden laskukaava on siis:

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

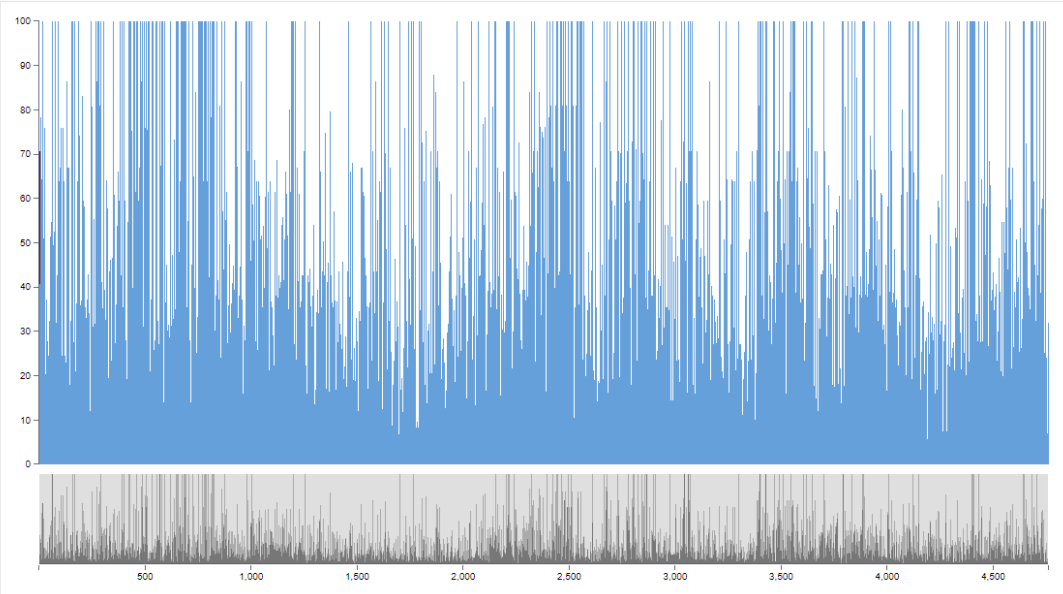
Kaava 3. BLEU-pisteiden laskuyhtälö. (Papineni et al. 2002: 315.)

Luonnolliselle kielelle on ominaista, että sama asia voidaan sanoa usealla eri tavalla. Tästä syystä lähtökielen lauseelle voi olla monta eri käännösvaihtoehtoa, joista kaikki ovat oikein. Käännösvaihtoehdoissa voi olla eri sanavalintoja tai niiden sanajärjestys voi olla erilainen. (Papineni et al. 2002: 312.) BLEU:ssa käytetään useita referenssikäännöksiä, jotta erilaisia sanavalintoja ei tulkittaisi virheeksi. BLEU asettaa kuitenkin hyvin vähän rajoituksia sille, kuinka n-grammivastaavuuksia määritetään eri referenssikäännöksistä. Vähäisistä rajoituksista johtuen BLEU sallii erittäin suuren määrän vaihtelua – jopa niin suuren, että sitä ei tulisi sallia, kun kyse on kääntämisestä. (Callison-Burch, Osborne & Koehn 2006: 251.) Vaihtelun ansiosta suuri määrä erilaisia ja eritasoisia käännöksiä voi saada samat BLEU-pisteet, ja pisteiden parantuminen ei aina johda käännöksen laadun parantumiseen, sillä BLEU ei välttämättä tunnista leksikaalista vaihtelua tai käännöksiä, joissa ei ole päällekkäisyyttä referenssikäännöksiin. (Callison-Burch, Osborne & Koehn 2006: 3–5.) Tästä syystä tutkielman testiaineisto pidettiin suhteellisen pienenä (n. 400 segmenttiä), jotta käännösten laatua olisi mahdollista arvioida myös manuaalisesti.

4 Tulokset

Tässä luvussa esitellään esikoulutetun konekääntimen uudelleenkoulutuksen tuloksia ja tarkastellaan tulosten taustalla olevia tekijöitä.

Konekääntimen uudelleenkouluttamisen jälkeen testiaineisto käännettiin uudelleen ja tuloksia vertailtiin. Tilden Interactive BLEU score evaluator -ohjelma laski esikoulutettujen konekäännösten BLEU-arvoksi 23,7 ennen uudelleenkoulutusta. Alla oleva kuvaaja havainnollistaa segmenttikohtaisia BLEU-arvoja niin, että x-akselilla ovat kaikki aineiston n. 5000 segmenttiä ja y-akselilla niiden BLEU-arvot.



Kuva 8. Segmenttikohtaiset BLEU-arvot ennen uudelleenkoulutusta. Laskettu Tilden Interactive BLEU score evaluator -ohjelmalla.

Kuvaajasta ilmenee, että vaikka suurin osa aineiston segmenteistä sijoittui 40 pisteen alapuolelle, sai osa segmenteistä silti hyvin korkeita pisteitä. Toisin sanoen esikoulutettu konekäännin osasi kääntää osan segmenteistä täysin ihmisen tekemän referenssikäännöksen mukaisesti ilman uudelleenkoulutusta. Suurin osa täydet BLEU-pisteet saaneista käännöksistä olivat odotetusti nominilausekkeita, mutta myös satunnaiset yksinkertaiset verbilausekkeet saivat korkeita pisteitä. Nomini- ja verbilausekkeita käsiteltiin tarkemmin luvussa 2.3.2. Alla olevassa taulukossa on esimerkkejä sekä nomini- että verbilausekkeista.

Lähtöteksti	Ihmiskäännös	Esikoulutettu MT	BLEU
Huppu on säädettävä.	The hood is adjustable.	The hood is adjustable.	100
Raglanhihat.	Raglan sleeves.	Raglan sleeves.	100
Erinomainen hengittävyys.	Excellent breathability.	Excellent breathability.	100
Kaksi tilavaa reisitaskua.	Two spacious thigh pockets	Two spacious thigh pockets	100
Laine on tehty pohjoismaisesta lohennahkasta.	Laine is made of Nordic salmon skin.	Laine is made of Nordic salmon skin.	100
Tämä vaate on vedenpitävä.	This garment is waterproof.	This garment is waterproof.	100
Pitkässä huivissa on viimeistelemättömät päädyt.	The long scarf has unfinished ends.	The long scarf has unfinished ends.	100

Taulukko 1. Esimerkkejä esikoulutetun konekääntimen tuloksista.

Kun esikoulutettu konekäännin uudelleenkoulutettiin tätä tutkielmaa varten kerätyllä aineistolla, nousi BLEU-arvo 13,2 pisteellä ja oli näin ollen 36,9. Suurin parannus tapahtui erikoisalan termeissä, jotka neuroverkkokonekäännin oppi tehokkaasti pienestä koulutusaineistosta huolimatta. Alla olevassa taulukossa on esitelty käännösesimerkkejä siten, että tarkastelun alla olevat sanat on lihavoitu. Taulukossa 1 MT tarkoittaa esikoulutettua konekäännintä ja 2 MT uudelleenkoulutettua konekäännintä.

Lähtöteksti	Ihmiskäännös	1 MT	1 BLEU	2 MT	2 BLEU
Ilmastoivat vetoketjut pohkeissa.	**Ventilation zippers** along the calves.	**Air-conditioned zippers** in calves.	9,65	**Ventilation zippers** along the calves	100

Monipuolinen ja **mukava pipo.**	A versatile and **comfortable beanie**.	very versatile and **nice hat.**	17,97	A versatile and **comfortable beanie**.	100
Lyhythihainen Luja-mekko on valmistettu **ryhdikkäästä puuvillatvillistä** ja sitä somistaa mustan ja luonnonvalkoisen sävyinen **Harha-kuosi**.	The short - sleeved Luja dress is made of **firm cotton twill** and decorated with the **Harha (illusion) pattern** in black and off white.	The short - sleeved Luja dress is made of stylish **cotton wool** and is adorned by a black and off-white **delusion.**	36,76	The short-sleeved Luja dress is made of **firm cotton twill** with the **Harha (Illusion) pattern** in black and off white.	86,0679
Kestävä ja lämmin **villatoppaus**.	With durable, warm **wool padding.**	A durable and warm **wool tap**	15,62	Durable, warm **wool padding.**	84,64817
Kolmisuuntaisesti säädettävä huppu, jossa säältä suojaava **vahvistettu lippa.**	3-way adjustable hood with **reinforced peak** for bad weather.	A three-way adjustable hood with a **reinforced cap** protecting against the weather.	28,92	3-way adjustable hood with **reinforced peak** for weather protection.	80,70557
GEL™-**kantapääpehmuste vaimentaa tärähdyksiä** tehokkaasti.	GEL™ **rear foot cushioning** for excellent **shock absorption**.	The geltm heel padding effectively **suppresses concussion**.	4,3	GEL™ **rear foot cushioning** for excellent **shock absorption.**	77,25506
Alanko-**tunika** on valmistettu **viskoosikrepistä** ja siinä on **V-pääntie**, **väljät vajaamittaiset hihat** sekä **helmassa sivuhalkiot**.	The Alanko **tunic** is made of **viscose crepe** and it has a **v-neck**, **loose cropped sleeves** and **side slits at the hemline**.	Alanko-**tunica** is made of viscous crepe and has **a v-head road**, **loose undersized sleeves** and **side hairs in the bosom.**	20,32	The Alanko **tunic** is made of **viscose creped** and it has a **V-neck**, **loose cropped sleeves** and **side slits.**	73,86387

Taulukko 2. Esimerkkejä erikoisalan termeistä ennen uudelleenkoulutusta ja sen jälkeen.

Termien osalta tulokset ovat hyvin positiivisia, sillä pienestä aineistosta huolimatta BLEU-arvo nousi merkittävästi sellaisissa segmenteissä, jotka koostuvat yksinkertaisista nominilausekkeista ja jotka sisältävät etuattribuutteja, jotka ovat suurimmassa osassa tapauksia tuotetta kuvaavia adjektiiveja. Uudelleenkoulutuksen myötä konekäännin oppi myös käyttämään haluttua varianttia, esim. *trousers* (vrt. pants) ja *jacket* (vrt. coat).

Suurin syy siihen, miksi BLEU-pisteet eivät nousseet 13,2 pistettä enempää, on uudelleenkouluttamisen myötä heikentynyt esikoulutetun konekääntimen laatu etenkin monimutkaisemmissa lauserakenteissa, jotka sisälsivät luovan kielen (ks. 2.2.2) elementtejä. Uudelleenkoulutuksen jälkeen konekäännin kadotti kontekstin monessa sellaisessa segmentissä, joiden kääntämisestä se oli suoriutunut hyvin tai keskinkertaisin BLEU-pistein ennen uudelleenkoulutusta. Alla olevassa taulukossa esimerkkikäännökset ovat huomattavasti parempilaatuisia ennen uudelleenkoulutusta.

Lähteteksti	**Ihmiskäännös**	**MT 1**	**1 BLEU**	**MT 2**	**2 BLEU**
Myös liikkumavapaus on tärkeää, kun tuote on aktiivisessa käytössä.	When a product is in active use, it is not irrelevant how it feels.	Freedom of movement is also important when the product is in active use.	25,92	It is also important to freedom of movement in active use.	11,88753
Vauhdikas meno ei ole ongelma, sillä kangas on erittäin hengittävä ja tuuletusvetoketjujen avulla ilmankiertoa voidaan tehostaa sykkeen noustessa.	Increasing the pace is not a problem as the fabric is highly breathable and the ventilation can be improved using ventilation zips when the heart rate starts to climb.	Speedy spending is not a problem, as the fabric is highly breathable and ventilation zippers make air circulation more efficient when the heart rate rises.	35,99	with this wearer, the fabric is not world you get a highly breathable and ventilation zips means that any air can be enhanced by autumn blend.	10,86127
Silence Pro -housut ovat täydellinen valinta, kun etsit vedenpitäviä	Silence Pro pants are perfect when you want a waterproof	Silence Pro pants are the perfect choice when looking	20,69	The waterproof outdoor pants are a perfect choice for all	9,23843

housuja, joissa on hyvä ilmanvaihto.	garment with ventilation.	for waterproof pants with good ventilation.		kinds of ventilation.	
Absidit voidaan kääriä kokonaan ylös teltan molemmilla puolilla, jolloin voit nauttia maisemista sisäteltasta käsin pitäen hyönteiset loitolla.	The vestibules can be entirely rolled up on both sides of the tent so you can enjoy the view from the inner tent while keeping insects at bay.	The azides can be wrapped up completely on both sides of the tent, allowing you to enjoy the landscape from the inner tent, keeping the insects away.	29,23	Asymmetric can be rolled up to this tent and that lets you enjoy the view of your insects.	7,487348

Taulukko 3. Esimerkkejä uudelleenkouluttamisen aiheuttamasta laadun heikkenemisestä ja käännösten lyhentymisestä.

Laadun heikentyminen johtuu uudelleenkoulutusaineiston kieleltään hyvin rajoittuneesta materiaalista ja lyhyistä virkkeistä (ks. 2.3.1). Konekäännin unohtaa uudelleenkoulutuksen myötä esikoulutuksessa oppimiaan asioita, vaikka esikoulutusaineisto onkin satoja kertoja suurempi kuin uudelleenkoulutuksessa käytetty aineisto. Lisäksi tuloksia analysoidessa ilmeni, että laadun heikentymisen ohella uudelleenkoulutus vaikutti käännettävien segmenttien pituuteen niin, että uudelleenkoulutuksen jälkeen käännökset olivat hieman lyhyempiä kuin ennen uudelleenkoulutusta ja poistoja oli enemmän, kuten taulukon 3 punaisella merkityissä esimerkeissä.

Tästä syystä uudelleenkoulutusta ei enää toistettu samalla aineistolla, vaikka aluksi koulutus oli tarkoitus toistaa useita kertoja sillä oletuksella, että koulutuksen toistamisella voitaisiin kompensoida aineiston pienuutta ja että BLEU-pisteet nousisivat jokaisen koulutuksen myötä. Koska BLEU-pisteet nousivat 13,2 pisteellä jo ensimmäisen koulutuksen jälkeen, mutta konekäännin unohti useassa segmentissä esikoulutuksessa oppimiaan ilmaisuja, päätettiin tällä aineistolla kouluttaa vain yhden kerran. Kuten luvussa 3.2 todettiin, samankaltaiseen tulokseen päätyivät myös esimerkiksi Freitag & Al-Onaizan (2016), joiden käännöslaatu parani BLEU-

pisteillä mitattuna saksa-englanti kieliparissa kahden koulutuskerran jälkeen ja kiina-englanti kieliparissa kuuden koulutuskerran jälkeen.

Ratkaisuna unohtamisongelmaan lisäsimme koulutusaineistoon n. 5000 sellaista segmenttiä, jotka olivat mukana myös esikoulutusmateriaalissa. Segmentit haettiin ja valittiin Korp-palvelun (ks. Borin, Forsberg & Roxendal 2012) avulla niin, että hakusanoina käytettiin tekstiilialalle tyypillisiä ilmaisuja. Lisäksi yhtenä valintakriteerinä vaikutti segmenttien pituus, sillä pidempien segmenttien ajateltiin vaikuttavan positiivisesti uudelleenkouluttamisen jälkeen tapahtuneeseen käännösten lyhentymiseen. Kaikki valitut segmentit olivat peräisin Opus-korpuksesta, eli niitä on käytetty myös konekääntimen esikouluttamiseen. Lisäämällä esikoulutusmateriaalissakin mukana olleita segmenttejä uudelleenkoulutusaineistoon pyrimme parantamaan sellaisten käännösten laatua, joihin uudelleenkoulutus vaikutti negatiivisesti. Esikoulutusmateriaalin hyödyntäminen auttoi, sillä BLEU-pisteet nousivat hieman alle pisteellä ja olivat näin ollen 37,7 kun testiaineisto käännettiin uudelleen. Alla olevassa taulukossa on esitelty kaksi esimerkkiä niin, että erot ihmiskäännöksen ja konekäännösten välillä on merkitty punaisella ja yhtäläisyydet vihreällä.

Esimerkki a	**BLEU**	Käännös
Ihmiskäännös	100.00	A great addition to every woman's wardrobe.
Uudelleenkoulutettu	28.32	This great addition to every garment.
Tuotekuvaukset + Korp	45.94	This is a great addition to every wardrobe.
Esimerkki b		Käännös
Ihmiskäännös	100.00	The pleasantly warm Torstai midlayer looks fabulous over a shirt.
Uudelleenkoulutettu	12.51	This warm Torstai midlayer is wonderful to put on and take off over a comfortable shirt.
Tuotekuvaukset + korp	28.92	The comfortable warm Torstai midlayer is ideal for wear over a shirt.

Taulukko 4. Esikoulutusmateriaalin lisäämisestä seurannut BLEU-pisteiden parantuminen yksittäisissä segmenteissä.

Kuten taulukosta 4 ilmenee, yksittäisten segmenttien BLEU-pisteitä pystyttiin parantamaan lisäämällä esikoulutusmateriaalia uudelleenkoulutusmateriaalin joukkoon. Testiaineiston BLEU-nousi kuitenkin vain hieman alle yhdellä pisteellä, eli kyse ei ole kovin merkittävästä parantumisesta. Lisättyjen segmenttien määrään nähden (n. 5000 segmenttiä) tulos on kuitenkin kannustava.

Koska esikoulutusmateriaalia lisäämällä ei saatu korjattua kaikkia uudelleenkoulutuksesta aiheutuneita ongelmia, on seuraava työvaihe muuttaa käytettyjä asetuksia niin, että ylikoulutukselta vältyttäisiin. Ylikoulutusta esiintyy kaikissa neuroverkkoarkkitehtuureissa: vaikka laatu paranisi aluksi, huononee se jälleen jossain vaiheessa koulutusta, kun vastaan tulee ennalta tuntemattomia esimerkkejä. Validointi on menetelmä, jota käytetään neuroverkon valvotussa koulutuksessa (eng. *supervised learning*) ylikoulutuksen havaitsemiseksi. Early stopping -tekniikalla pyritään estämään ylikouluttaminen, sillä early stopping pysäyttää käynnissä olevan koulutuksen heti, kun validointisetin virhelukema on edellisen validointisetin lukemaa korkeampi. Koulutusajoon käytetään näin ollen niitä painoja, jotka neuroverkolla oli käytössä ennen early stoppingia. (Prechelt 2000.)

Jatkotutkimuksessa työtilaa (eng. *workspace*) voidaan vähentää huomattavasti pienempien erien tekemiseksi, validointiväliä voidaan pienentää (validointi joka 500. tai 1000. erä) ja validointisettinä tulee käyttää kohtuullisen suurta validointijoukkoa (väh. 1000 lauseparia). Lisäksi early stopping voidaan määrittää niin, että ajo päättyy viiden validointivaiheen jälkeen, jos parannuksia ei ilmene. Validointisettiä käytetään siis esimerkkikohtaisten virheiden arviointiin viiden koulutusajon välein. Näitä asetuksia muuttamalla ylikoulutuksen riski pienenee, jolloin myös laadun voidaan olettaa paranevan ja BLEU-pisteiden nousevan.

5 Yhteenveto

Tutkielman tavoitteena oli tutkia tekstiilialan tuotekuvauksia ja niiden sopivuutta neuroverkkokonekääntämiseen. Teoriaosuudessa esiteltiin tuotekuvauksia, niiden rakennetta sekä kielellisiä piirteitä ja tarkasteltiin erilaisia tuotekuvauksia koskevia rajoituksia. Kyseessä on toistaiseksi hyvin vähän tutkittu aihe, joka tarjoaa varmasti runsaasti mahdollisuuksia jatkotutkimukseen. Tuotekuvausten lisäksi teoriaosuudessa keskityttiin neuroverkkokääntämiseen ja sen tällä hetkellä suosiossa olevien arkkitehtuurien ja teknologioiden esittelyyn, jotta tutkielman lähtökohtia ja tuloksia olisi helpompi ymmärtää ja tulkita.

Tutkielman tuloksena voidaan todeta, että esikoulutetun geneerisen neuroverkkokääntimen voi adaptoida erikoisalalle myös hyvin pienellä määrällä koulutusmateriaalia niin, että BLEU-pisteet nousevat jo heti ensimmäisen koulutusajon jälkeen. Tulos on rohkaiseva, sillä pienempiä rinnakkaiskorpuksia pystyy koostamaan myös itse sen sijaan, että käyttäisi jo olemassa olevaa korpusta – näin myös mahdollisten tutkimusaiheiden määrä kasvaa.

Kun esikoulutettu konekäännin uudelleenkoulutettiin tätä tutkielmaa varten kerätyllä aineistolla, nousi testikäännösten BLEU-arvo 13,2 pisteellä ja oli näin ollen 36,9. Suurin parannus tapahtui erikoisalan termeissä, jotka neuroverkkokonekäännin oppi tehokkaasti pienestä koulutusaineistosta huolimatta. Koulutusaineisto aiheutti kuitenkin myös ongelmia: tuotekuvausten hyvin rajoittunut kieli ja lyhyet virkkeet heikensivät konekääntimen laatua ja aiheuttivat käännössegmenttien lyhentymistä ja poistoja. Ongelmaa ei saatu kokonaan korjattua, mutta laatu parani hieman, kun uudelleenkoulutus toistettiin niin, että esikoulutusmateriaalissakin mukana olleita segmenttejä lisättiin uudelleenkoulutusaineistoon.

Vaikka tutkielman tulokset vastaavat muita aiheesta tehtyjä tutkimuksia ja tulosten luotettavuuteen on panostettu, tarjoavat tämän tutkielman tulokset vain hyvin rajoitetun katsauksen neuroverkkokonekääntimen adaptointiin erikoisalalle. Konekääntimen laatuun voi vaikuttaa uudelleenkoulutuksen jälkeen esimerkiksi yhdistämällä esikoulutetun- ja uudelleenkoulutetun mallin (Freitag & Al-Onaizan 2016) tai Ensemble-learningin avulla (ks. Freitag, Al-Onaizan & Sankaran 2017).

Tuotekuvaukset tarjoavat runsaasti jatkotutkimusmahdollisuuksia, ja erikoisalakoulutetulle neuroverkkokonekääntimelle on varmasti kysyntää alati kasvavilla markkinoilla, kun käännöksiä

toivotaan nopeilla aikataululla ja edullisin kustannuksin. Toivottavasti tämä tutkielma toimii pohjamateriaalina ja innoituksena jatkotutkimukselle.

Lähteet

Bahdanau, D., Cho, K. & Bengio, Y. 2016, Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473v7 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1409.0473.pdf (Haettu 15.2.2020).

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. 2016, Neural versus Phrase-Based Machine Translation Quality: a Case Study, Association for Computational Linguistics, Austin, Texas, s. 257–267.

Callison-Burch, C., Osborne, M. & Koehn, P. 2006, Re-evaluating the Role of Bleu in Machine Translation Research, Association for Computational Linguistics, Trento, Italy. Saatavilla: https://www.aclweb.org/anthology/E06-1032.pdf (Haettu 2.1.2020).

Calude, A. 2003, Machine translation of various text genres, Te Reo–the New Zealand Linguistic Society Journal, 46, s. 67–94. Saatavilla: https://www.researchgate.net/publication/228938192_Machine_translation_of_various_text_genres (Haettu 11.3.2020).

Chinea-Ríos, M., Peris, Á & Casacuberta, F. 2017, Adapting Neural Machine Translation with Parallel Synthetic Data, Association for Computational Linguistics, Copenhagen, Denmark, s. 138–147. Saatavilla: https://www.aclweb.org/anthology/W17-4714.pdf (Haettu 26.4.2020).

Cho, K., van Merrienboer, B., Bahdanau, D. & Bengio, Y. 2014b, On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259v2 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1409.1259.pdf (Haettu 4.3.2020).

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. 2014a, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1406.1078.pdf (Haettu 5.12.2020).

Cloud, A. 2018, *Self-Attention Mechanisms in Natural Language Processing*. Saatavilla: https://medium.com/@Alibaba_Cloud/self-attention-mechanisms-in-natural-language-processing-9f28315ff905 (Haettu 9.3.2020).

Crabbe, S. 2010, Controlled Languages for Technical Writing and Translation. Teoksessa *The Changing Face of Translation: Proceedings of the Ninth Annual Portsmouth Translation Conference Held on 7 November 2009*, toim. Kemble I., University of Portsmouth, s. 48–62.

Dannenberg, A. 2004, *Puhutun kielen segmentointi lausemaisiksi yksiköiksi*, Pro gradu -tutkielma, Helsingin yliopisto.

Domingo, M., Garcıa-Martınez, M., Helle, A., Casacuberta, F. & Herranz, M. 2018, How Much Does Tokenization Affect Neural Machine Translation? *arXiv:1812.08621v3 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1812.08621.pdf (Haettu 15.2.2020).

Forcada, M. 2017, Making sense of neural machine translation, Translation Spaces 6.2, s. 291–309. Saatavilla: https://www.dlsi.ua.es/~mlf/docum/forcada17j2.pdf (Haettu 6.2.2020).

Freitag, M. & Al-Onaizan, Y. 2016, Fast Domain Adaptation for Neural Machine Translation. *arXiv:1612.06897v1 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1612.06897.pdf (Haettu 26.4.2020).

Freitag, M., Al-Onaizan, Y. & Sankaran, B. 2017. Ensemble Distillation for Neural Machine Translation. *arXiv:1702.01802 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1702.01802.pdf (Haettu 30.4.2020).

Goldberg, Y. 2017, *Neural network methods in natural language processing*, Synthesis Lectures on Human Language Technologies 10.1, Morgan & Claypool, San Rafael, California.

Graves, A. 2012, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer-Verlag, Berlin Heidelberg. Saatavilla: https://www.cs.toronto.edu/~graves/preprint.pdf (Haettu 16.12.2019).

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. & Schmidhuber, J. 2015, LSTM: A Search Space Odyssey. *arXiv:1503.04069 [cs.NE]*. Saatavilla: https://arxiv.org/pdf/1503.04069.pdf (Haettu 15.2.2020).

Gross, A. 1992, Limitations of computers as translation tools, teoksessa *Computers in translation – a practical appraisal*, toim. Newton J., London, Routledge, s. 96–130.

Grygiel, M. 2017, *Cognitive approaches to specialist languages*, Cambridge Scholars Publishing, Newcastle upon Tyne.

Gröhn, A. 2019, *Suitability of Neural Machine Translation for Different Types of Texts: A Study on Potential Predictors,* Pro gradu -tutkielma, Helsingin yliopisto.

Haarala, R. 1981, *Sanastotyön opas*, Kotimaisten kielten tutkimuskeskus, Helsinki.

Helenius, T. 2016, Takaisinkytkeytyvät neuroverkot, teoksessa *Tietojenkäsittelytieteellisiä tutkielmia Talvi 2016*, toim. Mäkinen E., Tampereen yliopisto, s. 1–16.

Hochreiter S. & Schmidhuber J. 1997, Long Short-Term Memory, teoksessa *Neural Computation 9(8)*, s. 1735–1780.

Icepeak 2019, *Vedenpitävä vai vettä hylkivä?* Saatavilla: https://www.icepeak.fi/icepeak_fi/brandi/tuotetiedot (Haettu 2.2.2020).

Im, J. & Cho, S. 2017, Distance-based Self-Attention Network for Natural Language Inference. *arXiv:1712.02047 [cs.CL]*. Saatavilla*:* https://arxiv.org/pdf/1712.02047.pdf (Haettu 9.3.2020).

Isohookana, H. 2007, *Yrityksen markkinointiviestintä*, Talentum Media, Helsinki.

Jaaranen, K. 2000, *Kontrollerat språk vid maskinöversättning: en fallstudie av översättningsstrategier och subspråket i produktbeskrivningarna hos postorderföretaget Ellos Postimyynti Oy*. Pro gradu -tutkielma, Helsingin yliopisto.

Janzen, S. & Maass, W. 2008, Smart product description object (SPDO), Saarbrücken, Germany.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T. & Birch, A. 2018, Marian: Fast Neural Machine Translation in C++, Association for Computational Linguistics, Melbourne, Australia, s. 116–121. Saatavilla: https://arxiv.org/pdf/1804.00344.pdf (Haettu 2.12.2020).

Kaji, H. 1999, Controlled Languages for Machine Translation: State of the Art, MT SUMMIT VII: MT in the great translation era: proceedings of Machine Translation Summit VII, 13th-17th September 1999. Kent Ridge Digital Labs, Singapore. Saatavilla: https://pdfs.semanticscholar.org/e983/74435f3ff353e6748d2b1f7f6291fffae600.pdf (Haettu 3.10.2020).

Karlsson, F. 2008, *Yleinen kielitiede*, Uud. laitos. p., Gaudeamus Helsinki University Press, Helsinki.

Kielitoimiston ohjepankki: Pronominit. Saatavilla: http://www.kielitoimistonohjepankki.fi/haku/demonstratiivipronominit/ohje/549 (Haettu 4.2.2020).

Kim, I. & Kuljis, J. 2010, Applying Content Analysis to Web-based Content, CIT, 18. Saatavilla: https://www.researchgate.net/publication/220066206_Applying_Content_Analysis_to_Web-based_Content (Haettu 9.11.2019).

Koehn, P. & Knowles, R. 2017, Six Challenges for Neural Machine Translation, teoksessa *Proceedings of the First Workshop on Neural Machine Translation*, s. 28–39. Saatavilla: https://arxiv.org/pdf/1706.03872.pdf (Haettu 22.10.2019).

Kotler, P. & Keller, K. 2006, Marketing Management, Upper Saddle River, New Yersey.

Kuikka, L. 2009, Lehtimainoksen multimodaalisuus, teoksessa *Kielen piirteet ja tekstilajit: vaikuttavia valintoja tekstistä toiseen*, toim. Heikkinen V., Suomalaisen Kirjallisuuden Seura, Helsinki, s. 37–62.

Laenen, K. & Moens, M. 2019, Multimodal Neural Machine Translation of Fashion E-Commerce Descriptions, teoksessa *Fashion Communication in the Digital Age*, toim. Kalbaska, N., Sádaba, T., Cominelli, F., Cantoni, L., FACTUM 19 Fashion Communication Conference, Ascona, Switzerland, July 21-26, 2019, s. 46-57.

Laurén, C. 1993, *Fackspråk: form, innehåll, funktion*, Studentlitteratur, Lund.

Laurén, C. & Nordman, M. 1987, *Från kunskapens frukt till Babels torn: en bok om fackspråk*, Liber, Malmö.

Lehrberger, J. 1982, Automatic Translation and the Concept of Sublanguage, teoksessa *Sublanguage: studies of language in restricted semantic domains*, toim. J. Lehrberger & R. Kittredge, de Gruyter, Berlin, s. 82–106.

Lin, C. & Och, F.J. 2004, Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, Barcelona, Spain, s. 605–612.

Lipton, Z.C., Berkowitz, J. & Elkan, C. 2015, A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv:1506.00019 [cs.LG]*. Saatavilla: https://arxiv.org/pdf/1506.00019.pdf (Haettu 10.1.2020).

Loman, B. & Jörgensen, N. 1971, *Manual för analys och beskrivning av makrosyntagmer*, Studentlitteratur, Lund.

Luong, M., Pham, H. & Manning, C.D. 2015, Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1508.04025.pdf (Haettu 16.4.2020).

Malmelin, N. 2003, *Mainonnan lukutaito: mainonnan viestinnällistä luonnetta ymmärtämässä*. Väitöskirja, Helsingin yliopisto.

Manaswi, N.K. 2018, *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*, Apress, Berkeley, CA. Saatavilla: https://www.academia.edu/38704313/Deep_Learning_with_Applications_Using_Python_Chatbots_and_Face_Object_and_Speech_Recognition_With_TensorFlow_and_Keras_-_Navin_Kumar_Manaswi_Foreword_by_Tarry_Singh (Haettu 4.4.2020).

MarianNMT 2018, Saatavilla: https://marian-nmt.github.io/. (Haettu 29.4.2020).

Merisavo, M. 2008, *The interaction between digital marketing communication and customer loyalty*. Väitöskirja, Helsingin kauppakorkeakoulu.

Mitamura, T. 1999, Controlled Language for Multilingual Machine Translation, Machine Translation Summit VII, s. 46–52. Saatavilla: http://www.mt-archive.info/MTS-1999-Mitamura.pdf (Haettu 20.12.2020).

Mäkinen, M. 2019, *Domain adaptation: Retraining NMT with translation memories*. Pro gradu -tutkielma, Helsingin yliopisto.

Mustonen, A. 2001, *Mediapsykologia*, WSOY, Helsinki.

Niemikorpi, A. 1996, *Liikepostista tuikeilmaisimeen ja sulhasesta kuraenkeliin: erikoiskielten rakenteellisesta ja tyylillisestä vaihtelusta*, Vaasan yliopisto, Vaasa.

Nordman, M. 1994, *Minilekter: om de små textgenrernas språk*, Vaasan yliopisto, Vaasa.

Olah, C. 2015, *Understanding LSTM Networks*. Saatavilla: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (Haettu 15.2.2020).

Papineni, K., Roukos, S., Ward, T. & Zhu, W. 2002, BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, s. 311–318.

Pasanen, P. 2015, Terminologinen käsiteanalyysi asiatekstinkääntäjän työvälineenä, teoksessa *Käännetyt maailmat: johdatus käännösviestintään*, toim. Aaltonen S., Abdallah K. & Siponkoski N., Gaudeamus, Helsinki, s. 110–122.

Pedersen, D. 2014, Exploring the Concept of Transcreation - Transcreation as 'more than Translation? *Cultus Journal. The Journal of intercultural mediation and communication*, s. 57–71. Saatavilla: https://www.academia.edu/10238994/Exploring_the_concept_of_transcreation_transcreation_as_more_than_translation (Haettu 15.2.2020).

Pitkänen, K. 2008, *Suomi kasvitieteen kieleksi: Elias Lönnrot termistön kehittäjänä*. Väitöskirja, Suomalaisen Kirjallisuuden Seura, Helsinki.

Prechelt, L. 2000, Early Stopping - But When?, teoksessa *Neural Networks: Tricks of the Trade*, Springer Verlag, Berlin Heidelberg, s. 55–69.

Reiss, K., Roinila, P. & Vermeer, H.J. 1986, *Mitä kääntäminen on: teoriaa ja käytäntöä*, Gaudeamus, Helsinki.

Seljan, S. 2000, Sublanguage in Machine Translation, saatavilla: https://www.researchgate.net/publication/332396966_Sublanguage_in_Machine_Translation (Haettu 19.4.2020).

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M. & Liu, Y. 2015, Minimum Risk Training for Neural Machine Translation. *arXiv:1512.02433v3 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1512.02433.pdf (Haettu 1.4.2020).

Suomalainen, J. 2002, *Erikoiskielistä yleiskieleen – termeistä sanoiksi*. Saatavilla: https://www.kielikello.fi/-/erikoiskielista-yleiskieleen-termeista-sanoiksi (Haettu 28.1.2020).

Sutskever, I., Vinyals, O. & Le, Q. 2014, Sequence to Sequence Learning with Neural Networks, teoksessa *Advances in Neural Information Processing Systems*, toim. Ghahramani, Z., Welling, M., Cortes C., Lawrence N. D., Kilian Q. Weinberger, s. 3104–3112. Saatavilla: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (Haettu 2.1.2020).

Tiedemann, J. 2012, Parallel Data, Tools and Interfaces in OPUS, toim. N. Calzolari, K. Choukri, T. Declerck, et al, European Language Resources Association (ELRA), s. 2214–2218.

Toral, A. & Way, A. 2018, What Level of Quality can Neural Machine Translation Attain on Literary Text? *arXiv:1801.04962 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1801.04962.pdf (Haettu 4.5.2020).

Torresi, I. 2014, *Translating promotional and advertising texts*, 1st edition, Routledge, Taylor & Francis Group, London.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. 2017, Attention Is All You Need, teoksessa *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, Canada, USA. *arXiv:1706.03762v5 [cs.CL]*. Saatavilla: https://arxiv.org/pdf/1706.03762.pdf (Haettu 15.11.2020).

Vuokko, P. 1993, *Markkinointiviestintä*, WSOY, Helsinki.

Zhang, J. & Zong, C. 2015, Deep Neural Networks in Machine Translation: An Overview, *IEEE Intelligent Systems,* s. 16–25. Saatavilla: http://www.nlpr.ia.ac.cn/cip/ZongPublications/2015/IEEE-Zhang-8-5.pdf (Haettu 2.3.2020).

ENGLISH SUMMARY

University of Helsinki
Faculty of Arts
Master's Programme in Translation and Interpreting
Saara Salminen: Snowmobiles in the pants – Neural machine translation of e-commerce product descriptions
Master's Thesis 52 p., Appendix, English summary 7 p.
May 2020

---

1 Introduction

Like many industries, the textile and fashion industry is changing. Environmentally friendly and technical fabrics are developed in Finland at an increasing rate, and entering the international market is vital for many companies. Innovative models relating to sustainability and the circular economy have been taken into use.

However, entering international markets requires providing product information in other languages in addition to Finnish. Providing product information to consumers in their native language ensures they understand the features of the product and enables them to compare products. Translating product descriptions and websites to multiple languages is a laborious task for online stores, as many companies do not have appropriate processes in place for multilingual information management.

In addition to multilingual information management, quality control is challenged by the requirements and expectations set for product descriptions. Several translation strategies must be employed in translating product descriptions, since in addition to creativity, the translator is required to be familiar with the terminology of the industry. Errors in translations can lead to misleading consumers, for example, if a water-repellent coat is marketed as a waterproof one due to translation error. Consumer protection legislation forbids conveying false or misleading information in marketing, if the information is such that it could lead to consumers making a purchase decision or other decision related to the product that they would not have made without the provided erroneous information. Therefore, errors in product descriptions and their translations can lead to reclamations and added costs, when consumers return the jacket, marketed as

waterproof, due to is insufficient performance. The translation effort is made even more difficult by character restrictions, and guidelines and instructions relating to multilingual search engine optimisation.

Little research exists in the suitability of different text types to neural machine translation, but since the language of product descriptions is rather restricted and repetitive, it can be presumed that they would be very suitable for machine translation. On the other hand, product descriptions contain ambiguity and creative language, i.e. the kind of language that most often causes problems in machine translation. Thus, machine translation of product descriptions includes both the weaknesses and strengths of machine translation systems – which makes them an interesting area of study.

2 Theoretical Background

Before training a machine translation system, it is important to define the content, structure and linguistic features of product descriptions that separate textile industry product descriptions from all other text types. Little research exists on product descriptions; one target of this study is to provide a basis for the research of product descriptions by providing basic information on their function and linguistic characteristics.

2.1 Structure and linguistic features of product descriptions

A single text can combine linguistics characteristics of two or more text types (Reiss & Vermeer 1986, 116–117). Product descriptions combine features of informative, expressive and operative text types, since they simultaneously aim at providing information, creating positive mental images and associations, and at persuading consumers to purchase the product. However, the information conveyed with product descriptions is always carefully selected, as their primary function is a promotional one.

The language of product descriptions can be seen to include many characteristics of controlled language. However, product descriptions cannot be considered a controlled language, since, although based on natural languages, controlled languages are artificial. Product descriptions can be seen as a special language, since in addition to specific terminology, the descriptions contain specific stylistic and syntactic characteristics not common in standard language. On the other hand, product descriptions must be easily understandable in order to retain their function as

advertisements. Conveying industry-specific information to the general public requires simplifying, which means that the language used will also be closer to standard language (Laurén & Nordman 1987). Thus, establishing a clear definition for product descriptions is not as straightforward as it is for some other text types.

Product descriptions can be defined through their content. In addition to listing the key features of each product, the descriptions aim at piquing the reader's interest with various semantic means. Torresi (2014) has studied the translation of promotional material and other texts with a persuasive function. She concluded that in business-to-consumer marketing (b-to-c, B2C), texts always contain two types of characteristics: technical anchors and boost elements aimed at persuasion and sales promotion. This relates to the information-to-persuasion ratio: different genres have different ratios based on the elements they contain. Even the most technical texts can include elements that aim at persuasion and sales promotion, just as very creative and persuasive texts can contain technical facts. (Torresi 2014: p. 27.)

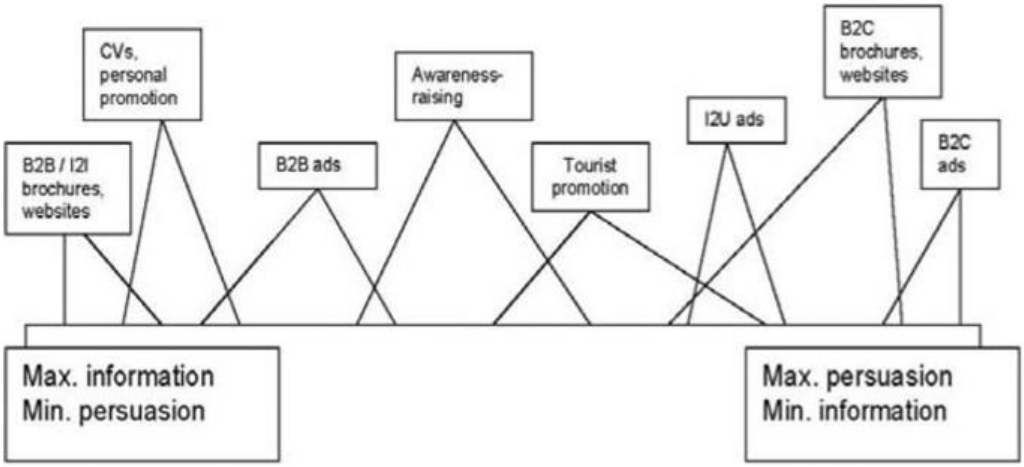


Figure 1. Information-to-persuasion ratios of different promotional and advertising text types (Torresi 2014: p. 28).

Product descriptions fall under the category of B2C promotional material on Torresi's continuum, as this category includes websites and brochures. Thus, the location of product descriptions on Torresi's continuum is among promotional texts that are more based on promotion and persuasion

than on facts. The large differences in the information-to-persuasion ratios of B2C marketing text, as seen in Figure 1, is also apparent in the fashion industry's product descriptions, since the ratio of providing technical information on the product and persuading the consumer varies according to the product being described (e.g. a functional sport jacket vs a simple polyester shirt). Brand image also has a significant effect on the information-to-persuasion ratio, as it provides guidelines for content production: a brand selling sporty outdoor clothing for the whole family will market its products in a different way than brands selling soft yoga clothing or products aimed at hardcore fitness enthusiasts.

For example, producing added value and differentiation are elements of product descriptions focusing on providing information and facts, and which contain multiple technical anchors. Product descriptions aim at adding value to the product by giving information on the features of the product that cannot be seen in images of the product and are not included in the feature list. There features particularly include all special materials and technologies (e.g. Dri-FIT, GORE-TEX) or information on the product being environmentally friendly.

Product descriptions that contain less facts focus more on persuading potential customers. This can be achieved with a variety of methods and strategies, but in the context of promotional material, Torresi especially employs the terms *creative language* and *emotional language* (2014: p. 121–128). Creative language includes metaphors, puns, neologisms, alliteration and onomatopoeia. Repetition and intertextuality are also important elements of creative language (Torresi, 2014: p. 121–124). Emotional language contains terms with either clearly positive or negative connotations, and clearly emotionally charged terms (e.g. dream, fantastic, magic). Instead of the products, focus can be on the consumer. In this case, first- and second-person personal pronouns are used. (Torresi, 2014: p. 128).

3. Material and Method

Training a neutral network for neural machine translation (NMT) requires large quantities of data. However, NMTs are considerably faster learners than their predecessors. NMT systems can usually be adapted to different domains rather well, but in these cases, translating text outside the NMTs domain will result in poor translation quality. Similarly, an NMT system trained with generic data, i.e. data that includes all kinds of text types from literature to subtitles and news

articles, will produce poor quality if used to translate texts that are very domain specific. (Koehn & Knowles, 2017: 1–2.) As a solution, NMT systems have been trained with generic training data first and domain-specific data after. This way, translation quality should improve and the NMT system should be able to handle domain-specific texts as well, even if the system is a generic model at its core (see e.g. Freitag & Al-Onaizan, 2016; Luong & Manning, 2015; Mäkinen, 2019). The same method was employed in this study as well.

Obtaining enough bilingual training material is one of the largest issues when it comes to training and developing NMT systems. Some promising solutions have been proposed as solutions. One is using synthetic data, i.e. selecting the most suitable training segments from a monolingual dataset that are then machine translated to create a parallel corpus (see e.g. Chinea-Ríos, Peris & Casacuberta, 2017: 138–145). In this thesis, the NMT was not trained with an existing corpus or with synthetic data. Instead, the training data consisted of bilingual product descriptions freely available online that were downloaded with Wget as html files and then aligned. This ensures that the retaining material does not include material that was also included in the training material of the pretrained NMT. The method also sheds light on how feasible it would be for a private person, such as a freelance translator, to compile a bilingual corpus with material downloaded from websites, for example. The compiled bilingual corpus can be used in other studies as well, such as in terminological studies.

The data includes product descriptions in Finnish and English from the following brands: Luhta, Icepeak, Rukka, Torstai, Marimekko, Minna Parikka, Fjällraven, Haglöfs, Lovia, Didriksons, Joutsen and Revolution Race. The source language of the product descriptions from the first four brands is Finnish. As for the rest, the source language is uncertain.

After downloading the Finnish and English texts, the html files were aligned with Memsource's alignment function, which can be used to open several alignment pairs for alignment simultaneously. After the files were aligned and cleaned, the resulting dataset of around 5000 translation segments was separated into two separate datasets, i.e. source segments and target segments (FI-EN). The Finnish segments were used as source language and the English material was used as reference for calculating the BLEU score.

After the Finnish source text was translated with the pretrained Marian NMT, a BLEU score was calculated for the translations, so that it could be compared with the BLEU score of the translations produced with the retrained NMT.

After this, the Marian NMT – pretrained by Professor Jörg Tiedemann with the data from the OPUS parallel corpus (see Tiedemann, 2012) – was retrained with the product description data, and the retrained NMT was used to translate the test data again. The BLEU score of the translations produced by the retrained model was calculated and compared against the BLEU score of the translations produced with the pretrained model.

4. Results

The test data was translated again after retraining the NMT and the BLEU scores were compared. Tilde's Interactive BLEU score evaluator gave the translations produced with the pretrained NMT a score of 23.7.

Although the score of most segments was below 40, there were segments that had a rather high score as well. The pretrained NMT was able to translate some segments identically when compared to the human-made reference translations even without the retraining. Most of the translations with high BLEU scores were noun phrases (as expected), but some simple verb phrases were given high scores as well.

When the pretrained NMT was retrained with the material collected for this study, the BLEU score increased by 13.2 points to 36.9. The most significant improvements were seen in domain-specific terminology, which the NMT learned very efficiently even with the small amount of training data.

Compared to the pretrained NMT, the retraining negatively affected the translation quality of complex sentences with high frequencies of the elements of creative language, which is the main reason why the BLEU score did not increase more than the achieved 13.2 points. The retrained NMT performed poorly in many segments that the pretrained NMT translated well or excellently (sometimes even with a BLEU score two times higher). The decrease in quality was caused by the very restricted language and short sentences of the retraining material. Retraining overrides some of the things NMTs learn from pretraining, even if the material used for pretraining was hundreds of times more extensive than the dataset used for retraining. This also caused the translations produced by the retrained NMT to be shorter.

5. Conclusions

In conclusion it can be stated that domain adaptation of a pretrained generic NMT is feasible even with a small set of training material and can lead to an increase in BLEU score even after the first training run. The results are encouraging, since compiling smaller parallel corpora is rather easy. Using self-compiled domain-specific corpora instead of existing larger corpora also opens up new avenues for further research.

When the pretrained NMT was retrained with the material collected for this study, the BLEU score increased by 13.2 points to 36.9. The most significant improvements were seen in domain-specific terminology, which the NMT learned very efficiently even with the small amount of training data. However, the training data caused some problems: the very restricted language and short sentences of the product descriptions reduced the quality of the machine translation, made the translated sentences shorter and added omissions. No complete solution was found to the problem, but adding some segments from the pretraining material to the retraining material and using the created dataset to train the model again was found to slightly improve quality.